

HUMAN LEARNING

in the school

John P. De Cecco

Readings in Educational Psychology



[PREFACE]

In recent years there has been an unfortunate void separating educational psychology and the psychology of learning. This book, intended as a text of supplementary readings for courses in educational psychology and the psychology of learning, takes a step toward bridging that void. The editor has attempted to provide here a balanced selection of articles that reflect contributions of both educational and experimental psychology to our knowledge of how children learn and how we can help them learn.

The emphasis throughout this book is on careful research. Fully half of the articles are direct research reports, and these are sufficiently detailed to provide the reader with a view of the experimental procedures used as well as the conclusions reached. Reports of research in the use of modern technology in the schools—television, films, language training, teaching machines, and programmed learning—are included, and there is a chapter devoted to communications. The remaining articles are summaries of large areas of research or interpretive discussions of major problems of human learning.

Editorial introductions to each chapter provide background for the research studies and point up parallel themes that recur in various readings. Both in the chapter introductions and in the introductions to the selections, the reader is encouraged to weigh the evidence and explore possible classroom applications.

As an aid to the instructor the editor has provided in the front of this book a cross-reference chart by which chapter assignments in current educational psychology textbooks can be coordinated with reading assignments in this volume.

Several criteria of selection derive from the book's major purpose

—to bridge the gap between the psychology of learning and educational psychology. First, all of the readings deal with some aspect of learning and apply directly or indirectly to school situations. Articles dealing solely with animal learning have been excluded; preference has been given to those on human learning in order to make easier the translation of laboratory research into testable hypotheses about classroom practice. Educational psychology is primarily a field of applied research, and, as such, the ultimate test of the usefulness of psychological research and theory must be their influence on classroom teaching and the improvement of classroom learning.

Second, for the beginning education student the reports of direct research that are included can serve as models of the kind of research that in the long run will prove most helpful to the classroom teacher. A common criticism of courses in educational psychology is that they are an organized body of bland, disembodied generalizations. In these research reports, however, the student can see for himself the factual basis for certain psychological concepts and principles. This exposure to direct research can also be a means of helping the student develop skill in inductive and conceptual thinking. The process of organizing hard-won information into a modest principle is a sobering experience: acquaintance with the content and method of psychological research in education may make one's generalizations more cautious and restricted.

Third, the general reports on large areas of research—for example, groups, communication, motivation—should give the student a broad view of problems associated with human learning and furnish a context for more specific research reports. In these general discussions, well-known researchers frequently comment on their own work and organize and review what has been accomplished and what remains to be done.

Because some distinction should be maintained between the study of development and the experimental study of classroom learning, and to avoid doing an injustice to both by attempting to treat them superficially in the same volume, the editor has, for the most part, omitted readings in human development. The chief contribution of information on development is to remind the teacher where he must set limits to his expectation for what he wants the child to learn.

Separate sections on the learning of attitudes and critical thinking have also been omitted.

Selections on the topics of intelligence and individual differences have been included in this text because the editor believes that the ability to learn and to handle abstractions is central to the learning process, and these articles are thus necessary to give a realistic picture of what the classroom teacher must be prepared to deal with.

Several readings in the area of mass media and audiovisual aids have been selected in order to stimulate the beginning education student to take a long-range view of what he and others may—or may not—be doing in their schools in the decades ahead. With programed learning and television in the vanguard, there are indications that the schools are in for much change. The student must, therefore, prepare for the curricula and schools of the future as well as for those of the present.

The editor gratefully acknowledges the authors and publishers who have generously consented to the reprinting of their articles. Special thanks is expressed to Dr. Norman Crowder of U.S. Industries, Inc., Dr. Felix Kopstein of Burroughs Corporation, and Dr. William Littell of San Francisco State College for permission to publish their articles for the first time; also to Dr. Bert Kersh of the Oregon State System of Higher Education and Dr. Merle Wittrock of the University of California, Los Angeles, for granting permission to publish articles prior to their appearance in the *Journal of Educational Psychology*; also to Dr. Robert Bostrom of Western Illinois University for revising his article for reprinting here.

The editor is deeply indebted to Professor Dale Harris of the Pennsylvania State University for the wise and consistent guidance and frank and constructive criticism he furnished throughout the preparation of the book. Special thanks is expressed to Professor Robert Gagné of Princeton University and the American Institute for Research, especially for his suggestions for the chapters on programed learning and human problem solving. Special thanks is owed Professor Warren Baller of the University of Nebraska for his careful reading of the manuscript and his valuable suggestions.

The editor wishes to thank his many colleagues at Michigan State University and San Francisco State College for their generous help

and encouragement. For inspiring an interest in the subject matter of this book the editor especially thanks Dr. Bernard Corman of Michigan State University and Dr. John Krumboltz of Stanford University. For greatly contributing to the editor's knowledge of educational technology and programmed learning, the editor thanks Dr. James Popham of the University of California, Los Angeles. For the opportunity to discuss and debate the many issues and problems in this book the editor thanks especially Dr. Dan Adler, Dr. Walcott Beatty, Dr. Sam Levine, Dr. Henry Lindgren, Dr. Jerome Podell, and Dr. Hilda Taba of San Francisco State College. For many patient weeks of typing and revision the editor expresses his thanks to Mr. Robert Flint, presently a doctoral student in psychology at the University of Minnesota.

J.P.D.

San Francisco
February 1963

[CONTENTS]

PREFACE	vii
CROSS-REFERENCE CHART	xvi
INTRODUCTION	1
CHAPTER 1	
<i>The Link between Laboratory and Classroom</i>	5
1 Walden II. B. F. SKINNER	10
2 The Science of Learning and the Technology of Educational Methods. ARTHUR W. MELTON	21
3 A Review of Factors in Learning Efficiency. R. M. GAGNÉ and R. C. BOLLES	30
4 The Teacher and the Improvement of Educational Practice. FREDERICK J. McDONALD	52
5 Viewpoints from Related Disciplines: Human Growth and Development. DAVID P. AUSUBEL	67
CHAPTER 2	
<i>Motivation: Does Curiosity Kill the Cat?</i>	81
1 Mice, Monkeys, Men, and Motives. HARRY F. HARLOW . . .	85
2 Manipulatory Motivation in Children. GLENN TERRELL . . .	98
3 Set Applied to Student Teaching. M. C. WITTRICK	107
4 Frustration and the Quality of Performance. IAN K. WATERHOUSE and IRVIN L. CHILD	117
5 Grades as Reinforcers in the Production of Attitude Change. ROBERT N. BOSTROM	129

CHAPTER 3

Programed Learning: The Ordering of Learning Situations 136

- 1 The Control of the Content of Conversation:
Reinforcement of Statements of Opinion.
WILLIAM S. VERPLANCK 140
- 2 Partial Reinforcement in a Gambling Situation.
DONALD J. LEWIS 156
- 3 Teaching Machines. *B. F. SKINNER* 164
- 4 The Rationale of Intrinsic Programming. *NORMAN A. CROWDER* 183
- 5 An Application of Recent Development in Psychology to the Teaching of German. *CHARLES B. FERSTER and STANLEY M. SAPON* 189
- 6 On the Relevance of Laboratory Investigation of Learning to Self-Instructional Programming. *THOMAS F. GILBERT* 199

CHAPTER 4

Human Problem Solving: To Find Out Is To Learn 209

- 1 Recent Research on Human Problem Solving.
CARL P. DUNCAN 212
- 2 The Act of Discovery. *JEROME S. BRUNER* 254
- 3 Inferential Behavior in Preschool Children. *HOWARD H. KENDLER and TRACY S. KENDLER* 271
- 4 The Motivating Effect of Learning by Directed Discovery.
BERT Y. KERSH 277
- 5 Experimental Studies in the Training of Originality.
I. MALTZMAN, S. SIMON, D. RASKIN, and L. LICHT . . 287
- 6 Pace of Presentation, Number of Trials, and Amount of Practice as Determiners of Learning. *FELIX F. KOPSTEIN* 311

CHAPTER 5

Communication: Words Still Count 323

- 1 How the Psychologist Views Communication.
IRVING LORGE 325
- 2 A Dispute about Reading. *ROGER BROWN* 338
- 3 Logic, Thinking, and Teaching. *B. OTHANEL SMITH* . . . 356
- 4 Verbal Context and the Recall of Meaningful Material.
GEORGE A. MILLER and JENNIFER A. SELFRIDGE . . 367

CHAPTER 6

The Mass Media: Films, Tapes, and Television 381

- 1 Stimulus-Response Psychology and Audio-Visual Education.
HOWARD H. KENDLER 384
- 2 ETV Research in the Decade Ahead.
E. G. SHERBURNE, Jr. 394
- 3 Comparative Effectiveness of Televised and Large Classes
and of Small Sections. LAURENCE SIEGEL and
F. G. MACOMBER 403

CHAPTER 7

Intelligence: Is Education Enough? 416

- 1 The Meaning and Assessment of Intelligence.
SIR CYRIL BURT 419
- 2 Three Faces of Intellect. J. P. GUILFORD 435
- 3 On the Growth of Intelligence. NANCY BAYLEY 456

CHAPTER 8

Individual Differences: All Men Are Created Unequal 479

- 1 Education and the Psychology of Individual Differences.
PHILIP E. VERNON 482
- 2 A Comparative Study of Achieving and Underachieving
High School Boys of High Intellectual Ability.
EDWARD FRANKEL 501
- 3 Social Class and Modes of Communication. LEONARD
SCHATZMAN and ANSELM STRAUSS 511

CHAPTER 9

Social Learning: Isn't It Always Better in a Group? 527

- 1 Groupness of the Group. IRVING LORGE 530
- 2 Group Problem Solving. DOROTHY I. MARQUART 537
- 3 Student-Centered vs. Instructor-Centered Instruction.
W. J. McKEACHIE 550
- 4 Attributed Social Power and Group Acceptance:
A Classroom Experimental Demonstration.
ALVIN ZANDER and ARTHUR R. COHEN 559

CHAPTER 10

<i>The Measurement of Learning: To Count Is To Number</i>	566
1 Educational Objectives and Curriculum Development. <i>BENJAMIN BLOOM, EDITOR</i>	569
2 Measurement in Relation to the Educational Process. <i>DOROTHY C. ADKINS</i>	575
3 Procedures for the Analysis of Classroom Tests. <i>ROBERT L. EBEL</i>	586
4 Must All Tests Be Valid? <i>ROBERT L. EBEL</i>	597
5 Further Comments on Test Validity. <i>WILLIAM W. LITTELL</i>	614
INDEX	627

TEXTS IN EDUCATIONAL PSYCHOLOGY

Listed in the Cross-Reference Chart (pp. xvi-xix)

- BIGGE, MORRIS L., and MAURICE P. HUNT. *Psychological Foundations of Education*. New York: Harper and Row, Publishers, 1962.
- BLAIR, GLENN M., R. STEWARD JONES, and RAY H. SIMPSON. *Educational Psychology*, rev. ed. New York: The Macmillan Company, 1962.
- COLE, LAWRENCE E., and WILLIAM F. BRUCE. *Educational Psychology*, rev. ed. New York: Harcourt, Brace & World, Inc., 1958.
- CRONBACH, LEE J. *Educational Psychology*, rev. ed. New York: Harcourt, Brace & World, 1963.
- FRANDSEN, ARDEN N. *Educational Psychology: The Principles of Learning in Teaching*. New York: McGraw-Hill Book Company, 1961.
- GARRISON, KARL C., and GRAY, J. STANLEY. *Educational Psychology*. New York: Appleton-Century-Crofts, 1955.
- KLAUSMEIER, HERBERT J. *Learning and Human Abilities: Educational Psychology*. New York: Harper and Row, Publishers, 1961.
- LINDGREN, HENRY CLAY. *Educational Psychology in the Classroom*, rev. ed. New York: John Wiley and Sons, Inc., 1962.
- MCDONALD, FREDERICK J. *Educational Psychology*. San Francisco: Wadsworth Publishing Company, Inc., 1959.
- MORSE, WILLIAM C., and G. MAX WINGO. *Psychology and Teaching*, rev. ed. Chicago: Scott, Foresman and Company, 1962.
- MOULY, GEORGE J. *Psychology for Effective Teaching*. New York: Holt, Rinehart and Winston, Inc., 1960.
- PRESSEY, SIDNEY L., FRANCIS P. ROBINSON, and JOHN E. HORROCKS. *Psychology in Education*, rev. ed. New York: Harper and Row, Publishers, 1959.
- SAWREY, JAMES M., and CHARLES W. TELFORD. *Educational Psychology*. Boston: Allyn and Bacon, Inc., 1958.
- SKINNER, CHARLES E. *Educational Psychology*, rev. ed. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1959.
- STEPHENS, J. M. *Educational Psychology: The Study of Educational Growth*, rev. ed. New York: Holt, Rinehart and Winston, Inc., 1956.
- THOMPSON, GEORGE C., ERIC F. GARDNER, and FRANCIS J. DIVESTA. *Educational Psychology*. New York: Appleton-Century-Crofts, 1959.
- TOWNSEND, EDWARD ARTHUR, and PAUL J. BURKE. *Learning for Teachers*. New York: The Macmillan Company, 1962.
- TROW, WILLIAM CLARK. *Psychology in Teaching and Learning*, rev. ed. Boston: Houghton Mifflin Company, 1960.

CROSS-REFERENCE CHART FOR USE WITH CURRENT

CHAPTER NUMBER IN TEXTBOOKS	BIGGE, HUNT (1962)	BLAIR, JONES, SIMPSON (Rev. 1962)	COLE, BRUCE (Rev. 1958)	CRONBACH (Rev. 1963)
1	I 1, 2, 3	I 1, 2, 4, 5		I 1, 2, 3, 4
2	I 1, 5	I 5		I 1; IX 1, 4
3	I 5	VIII 1, 2, 3		I 3
4	I 1	VIII 1, 2, 3	VI 1, 2, 3	II 1, 2
5	VII 1, 2, 3; VIII 1, 2, 3; X 4, 5	I 3	I 1	VIII 1, 2, 3
6		VIII 1, 2, 3; V 2	II 1, 2, 3, 4, 5	I 5; II 1
7	I 5	II 1, 2, 3, 4, 5; III 1, 2		V 2; VII 1, 2, 3
8		II 3, 5; IX 1, 4	I 2, 3; III 6	VII 1, 2; VIII 1; X 4, 5
9	II 1, 2	III 3, 4, 5, 6; IV 1, 2, 3, 4, 5	III 1, 2, 3, 5; IV 5, 6	IV 6; V 2
10	I 3; III 1, 2, 6; VI 1	IV 2, 3, 4	II 3, 5; III 1, 2	IV 1, 2, 3, 4, 5
11	I 3; III 1, 2; VI 1	IX 1, 2, 3, 4	IV 1, 2, 4	IV 1, 2, 3, 4, 5
12		III 3, 4	IV 1, 2, 3, 4	III 1, 2, 3, 4, 5, 6; V 1, 2, 3, 4
13	III 1, 2, 3, 4, 5, 6	II 4	IV 5	IX 1, 4
14	IV 1, 2, 3, 4, 5			II 1, 2, 3, 4, 5; III 1, 2
15	II 3, 4, 5; III 3, 4, 5; IV 2, 3, 4; VI 1, 2, 3			IX 1, 2, 3, 4
16	VII 1, 2, 3	I 5		X 1, 2, 3, 4, 5
17	IV 2, 3	X 1, 2, 3, 4, 5		III 4
18	IV 1, 2; IX 2	X 4, 5		
19	IX 1, 2, 3			
20				
21				
22		II 3		
23				
24				
25				

Roman numerals refer to chapter numbers in this book of readings. Arabic numerals refer to readings within the chapters.

TEXTBOOKS IN EDUCATIONAL PSYCHOLOGY

[illegible]

(Cont.)

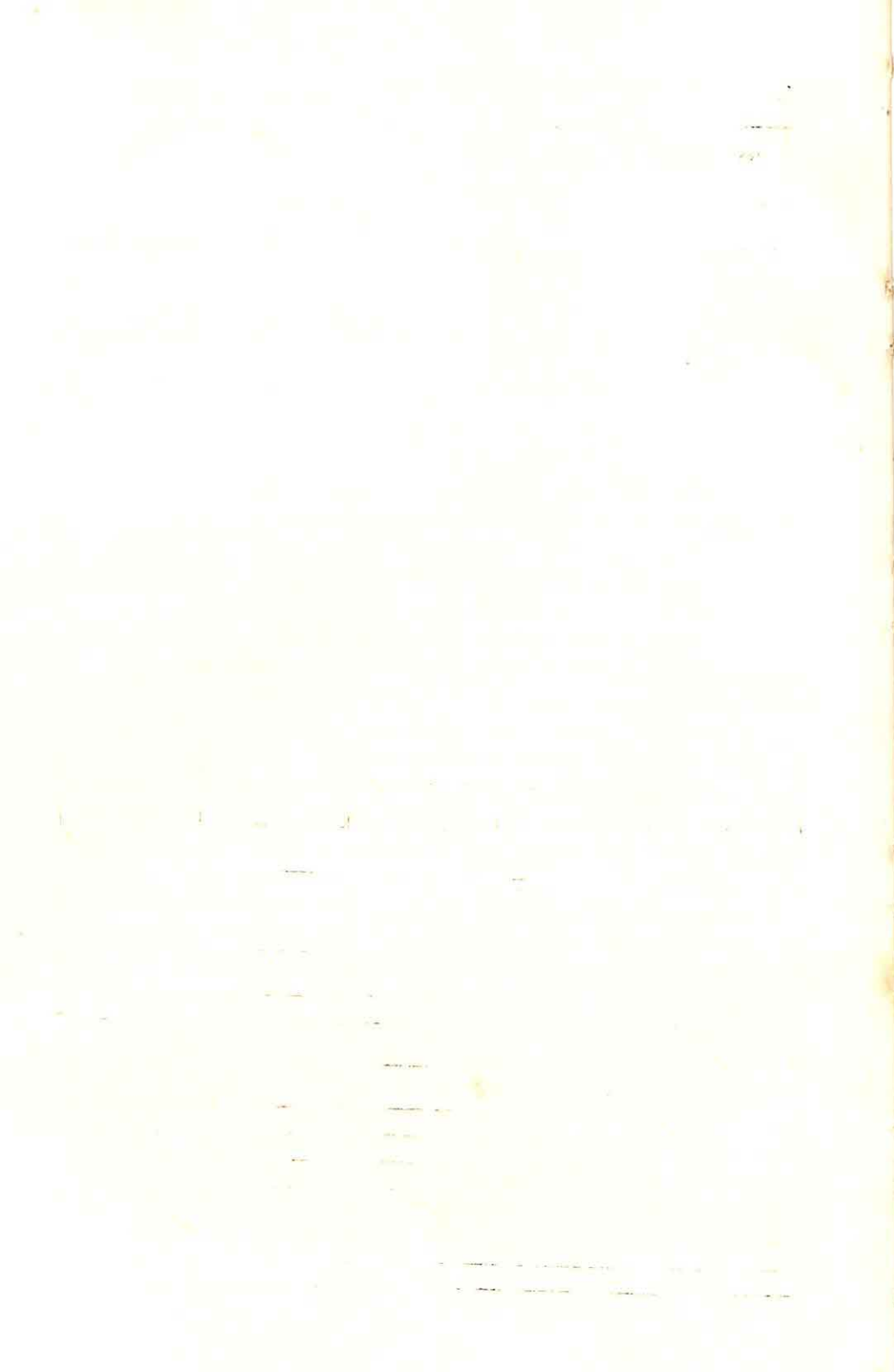
CROSS-REFERENCE CHART FOR USE WITH

MORSE, WINGO (Rev. 1962)	MOULY (1960)	PRESSEY, ROBINSON, HOBBOCKS (Rev. 1959)	SAWREY, TELFORD (1958)	SKINNER (ed.) (1959)
I 2, 3, 4	I 1, 2, 3	I 1, 2, 3, 4, 5	I 2, 3, 4, 5	I 1, 2, 3, 4
I 1; III 3, 6	II 1, 2, 3, 4, 5	I 5	II 1, 2, 3, 4, 5	I 3, 4
I 5	I 5; IV 3	VII 1, 2, 3	I 3, 4; III 1, 2; IV 1, 2	
II 1, 2, 5		II 1, 2, 3, 5	II 3; III 1, 2; IV 2; VI 1, 2, 3	I 5
II 1, 2, 4	I 1; II 4; IV 4	I 1; II 1, 4	IV 2, 3, 4, 5, 6; V 1, 2, 3	
	IX 1, 2, 3, 4		X 1, 2, 3, 4, 5	
III 3, 4, 5, 6; VI 1, 2, 3	VII 1, 2, 3	VIII 1, 2, 3	I 5	I 5; VII 1
IV 1, 2, 3, 4, 5, 6; V 4	I 3, 4; III 1, 2; IV 1	I 3, 4		
IV 1, 2, 3, 4, 6; V 2	III 3, 4, 5, 6; IV 2, 3, 4, 5, 6	IV 1, 2, 3, 4, 5, 6	VIII 1, 2, 3	VII 1, 2, 3
IV 2, 4	II 1, 2, 3, 5	IX 1, 2, 3, 4	II 1, 2, 3, 5	I 1; II 4
VII 1, 2, 3; VIII 1, 2, 3; X 2, 4, 5	IV 2, 4, 6	III 1, 2, 3, 4, 5; VI 1, 2, 3	I 1; II 4	VIII 3; IX 1, 3, 4
II 1, 2, 3, 4, 5	V 2	V 2	VIII 1, 2, 3	II 1, 2, 3, 4, 5
IX 1, 2, 3, 4	II 3; VIII 3; IX 1, 3, 4	X 1, 2, 3, 4, 5	IX 1, 2, 3, 4	IX 1
II 4	X 1, 2, 3, 4, 5	VIII 1	II 1, 4; III 3, 6	I 3, 4; III 1, 2
VIII 1, 2, 3	VIII 1, 2, 3			I 3; III 1, 2, 3
X 1, 2, 3, 4, 5	I 1; IX 1, 3, 4			II 1, 2, 3, 4, 5
	I 5	II 4		IV 1, 2, 3, 4, 5
	I 2, 4; III 3	I 4		IV 1, 2, 3, 4, 5, 6
				IV 1, 2, 3, 4, 5; V 1, 2, 3, 4
				III 3; V 2
				VIII 1, 2, 3
				I 2, 4
				VII 1, 2, 3; X 2, 4, 5
				X 1, 2, 3, 4, 5
				VIII 1, 2, 3

Roman numerals refer to chapter numbers in this book of readings. Arabic numerals refer to readings within the chapters.

CURRENT TEXTBOOKS IN EDUCATIONAL PSYCHOLOGY

STEPHENS (Rev. 1956)	THOMPSON, GARDNER, DIVESTA (1959)	TOWNSEND, BURKE (1962)	TROW (1960)	CHAPTER NUMBER IN TEXTBOOKS
I 1, 2	I 5	I 1, 2, 3, 4, 5	I 2, 3, 4, 5	1
I 4, 5	I 1	II 1, 2, 3, 4, 5	I 1; II 3, 5; IX 1, 3	2
	II 4	VII 1, 2, 3; VIII 1, 2, 3	III 3, 4, 5, 6	3
	VIII 1, 2, 3	IX 1, 2, 3, 4	II 1, 2, 4	4
VII 1, 2, 3	X 1, 2, 3	IX 3, 4; VI 1, 2, 3	VIII 1, 2, 3	5
VIII 1, 2, 3	VII 1, 2, 3	II 3, 4, 5	VII 1, 2, 3	6
VIII 1, 2, 3	II 1, 2, 5	I 3, 4	VIII 3	7
I 3	I 3	IV 1, 2, 3, 6		8
	IX 1, 2, 3, 4	IV 1, 2, 3, 6		9
		III 3, 4, 5; IV 2, 3, 4	III 1, 2, 3, 4	10
III 1, 2, 3, 4, 5, 6	I 2, 3; III 1, 2, 3, 4, 5	IV 2, 3, 4, 5	III 1, 2; V 1, 2, 3, 4	11
IV 1, 2, 3, 4; IX 2	II 1, 2, 3, 4, 5	III 1, 2, 3, 4, 5, 6	IV 1, 2, 3, 4, 5, 6	12
IV 4, 6; V 1, 3, 4; VI 1, 2, 3	IV 1, 2, 3, 4, 5	IV 1, 2, 5	IX 1, 2, 3, 4	13
IV 1, 2, 3, 4, 5, 6	IV 1, 2, 3, 4, 5		IX 3, 4	14
IX 4	IV 1, 2, 3, 4, 5; VI 1, 2, 3	IV 6	I 4	15
II 1, 2; VIII 1, 2, 3	II 3, 4; IX 1, 3, 4	IV 1, 2; V 3; IX 2		16
I 1		V 1, 2, 3, 4		17
IX 1, 2, 3, 4		X 1, 2, 3, 4, 5		18
				19
				20
				21
				22
				23
				24
				25



*Why
Education
and Psychology*

A major objective of this collection of readings is to explore the relationship of psychology to classroom education, especially in the light of contemporary developments in both fields, and to assess the contributions that one may make to the other in the future.

The central concern of both psychology and education is the study of man: They are both behavioral disciplines concerned with how man learns to solve the array of problems that constantly beset him—how he adapts, remembers, and forgets; his motives and feelings; his most noble aspirations and his most abject fears. They are both interested in his sensory perception of the physical world. Finally, in addition to their common interest in the similarities of men, both education and psychology are interested in how one man differs from another.

Another element common to both psychology and education is the use of the scientific method of inquiry. The invention of a method for posing a question and a hypothetical answer to it, and then *testing* that answer, has been an outstanding contribution of western civilization to the intellectual development of man. The conclusions of any inquiry in education or psychology must rest on the weight of evidence furnished by careful investigation; in neither field is tradition, authority, or personal intuition a sufficient basis for claims to knowledge. Of the behavioral sciences, psychology has been most rigorous in its use of the scientific method. Therefore, education, which is interested in many of the problems of psychology, has often looked to psychology for guidance in conducting its empirical in-

vestigations and for principles upon which to base its professional practice.

In order to show the education student the relationship between the study of education and the science of psychology, the articles presented here are largely experimental studies which not only present new evidence but also distinguish between competent and doubtful evidence. Since the results of scientific inquiries are most often presented in numerical data, it is necessary for the student to become familiar with sampling procedure, experimental design, and statistical analysis. But the mathematical aspects of scientific inquiry in education and psychology should in no way discourage the student in his attempts to understand the articles. What may initially appear to be somewhat difficult most often becomes progressively easier and more rewarding. Where assistance seems necessary in understanding the articles, interpretive comments have been added.

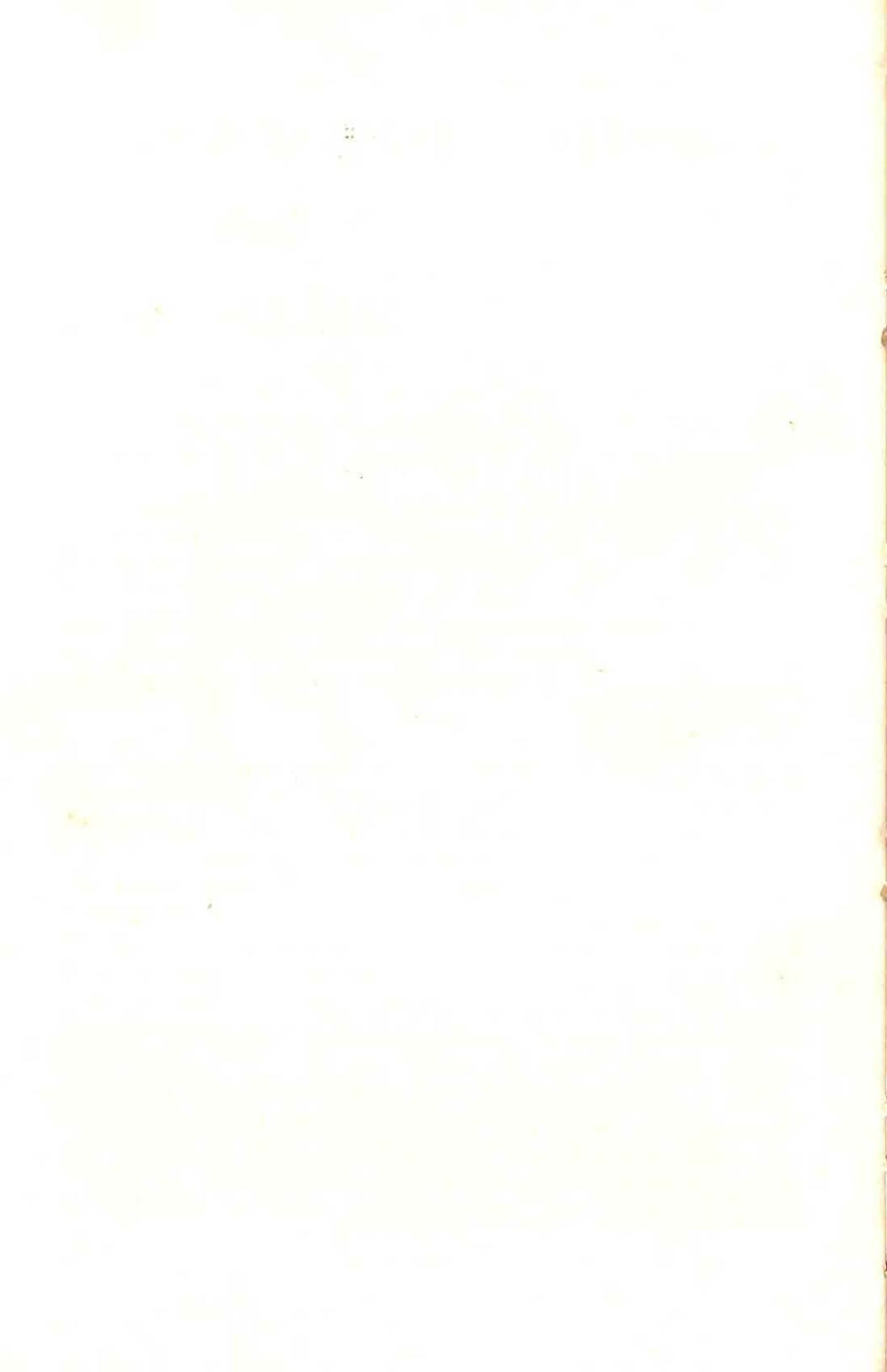
Although psychology and education share the characteristics that we have mentioned, there is one important difference between them. Psychology is a disinterested inquiry into human behavior, with its sole purpose being prediction and explanation. The usefulness of this knowledge, if any, is not a major concern of many research psychologists. Research in education, however, has as its major goal the improvement of the educative process. Given particular educational objectives, the educational researcher attempts to find those procedures and techniques that will most efficiently achieve them. Thus research in education is more programmatic than theoretical.

It has been said, with much aptness, that education stands in relation to psychology as engineering to the physical sciences. The difficulty with this analogy is that it may conceal the fact that many of the problems which psychologists study have little or no relevance to classroom learning. In the past the schools have sometimes borrowed wholesale from both clinical and experimental psychology. When the borrowed concepts and principles were inappropriate, the adaptation was made by converting the classroom into a clinic or laboratory. The alternative, of course, is to test in the classroom the principles taken from psychology to find out whether they assist in developing better methods for achieving educational objectives.

Another major objective of this book of readings is to encourage the education student to think about his future role in our society,

whether as behavior model or intellectual leader. In fact, it is the American teacher who can supply an attractive and genuine intellectual model for the American student to imitate, by patiently and skillfully gathering the information he needs to arrive at warranted conclusions, by knowing how to *use* the information he gathers, but by still realizing the limitations and tentative nature of any general statements he may make. Creating such an awareness of this aspect of his responsibility as a teacher is an aim of this book. In a course in educational psychology the student teacher can learn the methods and concepts of psychology which will enable him to think analytically about educational practice in the classroom as well as to plan classroom activities which will help his students to learn successfully. Such analytical and organizational activity on the part of the teacher rightfully confers upon him a professional status.

The way in which a person acquires knowledge determines whether or not he will retain or forget it, as well as whether he will later be able to apply what he has learned. As part of a course in educational psychology, this book of readings is intended to demonstrate to the student who is becoming a teacher how the principles of psychology will help him to organize his classroom practices so that his pupils will learn what he wishes them to learn.



[CHAPTER 1]

The Link between Laboratory and Classroom

The readings in Chapter 1, because of their broad, theoretical nature, may prove to be the most difficult in the book. The difficulty stems at least partly from the fact that, until recently, no one has thought systematically about the relationship between education and experimental psychology. In fact, many educators and psychologists never assumed there was one. Consequently, the student can find some excitement in the fact that, in this area, he is not far behind the "experts."

Any investigation of the relationship between psychology and education is difficult. It necessitates the raising of some thorny questions about the nature of knowledge and "truth," the nature of the scientific method, the present "structure" of psychological knowledge (especially psychological theory about learning), the nature of education, and the translation of the psychological laboratory's knowledge into the classroom's educational research and practice. Each of these questions is raised and discussed again and again in later chapters, and in more specific contexts, in the area of motivation for example. Only at the end of the course, could the instructor expect that the students' tentative answers to these questions would be fairly sophisticated.

The student probably has some stereotype of psychologists. The experimental, "rat" psychologist is supposedly uninterested in, if not hostile to, education. The educational psychologist, on the other hand, may be less misrepresented—if only because he may be quite unknown. The student may be surprised to find out that, although

both experimental and educational psychologists belong to the same field of study, few are familiar with research in both fields. Frequently they associate themselves with different educational currents, the experimentalist with the tradition of the liberal arts, the educationist with the front-line battles of the college of education.

To understand the dissociation of the two fields of psychology one must consider an academic tradition which, often more ritualistically than realistically, tries to maintain a division between theory, or the quest for knowledge alone, and practice, or the desire to do something better than we are doing it now. However, as Melton explains (pp. 27-28), pure and applied research belong to the same continuum and cannot be rigidly categorized. Gagné and Bolles have observed (pp. 31-32) that the experimentalist often studies learning apart from any of its practical implications—often using the kinds of materials which teachers never ask students to learn (for example, lists of nonsense syllables)—and that he might be as curious about learning losses—how things get “unlearned” (something which teachers dread but rarely try deliberately to foster)—as he is about learning gains. The educational psychologist must begin with the conditions, objectives, and materials of the school. He would like to know more about conditions which avoid losses in learning as well as about time spent in learning. The fact is that the two areas of research are not mutually exclusive. The separation is even more surprising to discover when one realizes that both fields use the same principles of research design and the same statistical methods for gathering and analyzing data. ✓

This separation has had disadvantages for both types of psychologists. It has meant that the experimental psychologist rarely reads the education journals or enters a classroom, other than his own, to discover new, practical research problems which could extend his knowledge about learning, especially human learning. Much of his research has been confined to animal learning. About him, Melton has stated:

To me, one obvious criticism of what has happened in the last twenty-five years is the domination of theories of learning by the rat. . . . To concentrate so much energy and scientific elegance on an animal subject with limited verbal capacities seems to be paying a high price for ease in procurement, maintainability, and reproduction of kind.

R | . . . In particular, it is my hope and belief that children of all ages will become the preferred subjects of theory and experiment in learning and problem solving.*

✓ These courageous words were spoken at an annual meeting of the American Psychological Association a few years ago. Now several leading experimental psychologists have turned their attention to educational problems, contributing their knowledge, skills, and zeal for reform. One of the most prominent has been B. F. Skinner (see pp. 10-20 and pp. 164-182), who is a leader in the development of teaching machines and programmed learning, both of which may revolutionize educational practice in the decade ahead. Another is ✓ Jerome S. Bruner (see pp. 254-270), who is investigating the implications for teaching which have come from his and other's research on how we form ideas (concepts). In some of the leading educational journals one may now find articles by such experimentalists as Postman, Kendler (see pp. 384-393), Underwood, Spence, and Deese. For the most part these articles are general discussions of the research on learning and suggestions for experimentation in education. The actual experimentation undoubtedly will have to be done by a type of individual presently in very short supply: the experimentally-trained educational psychologist who knows about the science of learning. It is conceivable that some education students, after they become teachers, may decide to secure the training necessary to do this research.

For the many educational psychologists of the past, or perhaps more generally, for the educational researchers, the schism has often resulted in research of a nonexperimental nature and of very narrow scope, often limited to studies of particular programs in particular schools and yielding few generalizations capable of application elsewhere. Such research has become notorious. It has often been a gigantic empirical undertaking in which tons of data are amassed, thousands of students tested, and more thousands of dollars spent. Only when Samson rests does someone ask what hypotheses were being tested: What did the researchers expect the program to ac-

* "Present Accomplishments and Future Trends in Problem Solving and Learning Theory," *American Psychologists*, 11 (1956), 278-281.

comply and what was the theoretical basis for expecting anything desirable to happen? And no one has the answer. Frequently financial support for research has been offered by private foundations before anyone has a problem or even a design for studying one. A research staff is quickly assembled, the project launched, the money consumed. The bustling about raises many hopes, which later dissolve as the staff dwindles away along with the money. Those who continue to hear about the project or maintain an association with it hope for its quiet burial. Unfortunately this fiasco is repeated again and again, because burials of this sort are so easily arranged that no one systematically inquires into what went wrong and how it could be avoided the next time. In all fairness to the educational researcher, it should be added that because he must conduct his research in a school setting, with all the restrictions this implies, his research has been limited to very descriptive, factual inquiries. However, because he has often been operating without the use of the concepts and principles of the learning psychologist, he has had no basis for constructing hypotheses of an experimental nature. It is as if the educational psychologist possesses all of the important problems of learning in our schools, while the experimentalist, with many skills and some theory, has a limited range of laboratory problems to work on.

All the materials for healing the schism are now present. First of all, the dichotomy between pure and applied research has always been particularly artificial for psychologists because they are linked by choice and career to the understanding and welfare of their fellow man. In the experimental psychologist, as he assembles the new apparatus which will force some unwary Norway rat into some incidental learning (that is, learning which the rat never intended), the zeal for probabilistic Truth may burn more brightly than the zeal to save Mankind, but because the truth he seeks concerns, in some very intimate way, human behavior, the distinction is more poetic than real. In the educational psychologist the desire to save the present and future generations from inefficient educational practices may be more intense than his desire to build a science of education, but because the reform he seeks must ultimately come from careful scientific investigation, this distinction is one which also lacks reality.

The Relationship of Readings in Chapter 1

The readings in the first chapter, as indicated above, are an attempt to describe the relationship of psychology to teaching. For the first reading the editor has chosen a selection from Skinner's *Walden Two*, the description of a psychologist's Utopia. Here it is my purpose to give the student a fairly concrete illustration of what it might mean if we really began to apply psychological knowledge systematically to the solution of social and educational problems. *Walden Two* is a palatable *Brave New World*. The article by Melton, a theoretical discussion of the relationship between psychology and education, outlines the contributions psychology could make to education if certain obstacles were removed and considers the profound alterations in educational practice which these contributions would make. The third reading, by Gagné and Bolles, sheds more light on this relationship; it gives a detailed outline of the conditions of the learning situation which the laboratory psychologist investigates and describes how these conditions could now be translated into educational practice as conditions to be manipulated by the teacher. The reading by McDonald is a careful explanation of how the teacher can act as an experimental psychologist in his own classroom, by employing psychological concepts in making his decisions in teaching and by maintaining a critical attitude toward what is considered "obvious." The last reading, by Ausubel, is a warning. The relationship which has existed between education and developmental psychology has resulted in harmful educational practices. The reading is a reminder that teachers cannot borrow wholesale from psychology and that a process of translation (through research) from laboratory to classroom is essential.

B. F. SKINNER
Harvard University

Walden Two *

Every book, including a book of readings, ought to have vision, although not that alone. It is not at all unlikely that on some warm spring day the education student will put his work aside, not for social dalliance, but because he wonders what difference he and others can make in our society. Because their roles prevent them from thinking of each other as basically human, the student may not realize that his professors sometimes do not find the ivory tower meaningful and that they also put their books and mazes aside to wonder what influence their work and education ultimately have on society.

Since fact cannot answer this question we often resort to fantasy. *Walden Two* is such a fantasy, created by B. F. Skinner, an experimental psychologist whose name is connected with learning theory (pp. 138-139) and teaching machines (pp. 164-182). It is a description of a society entirely formed by the methods and theory of experimental psychology. The following excerpt from the novel describes the education of the emotion of the children who grow up in Walden Two. It takes the form of a conversation between Frazier, the founder of Walden Two, and two university professors. One of these is Augustine Castle, a professor of philosophy, who is described earlier in the book as rather stout because of his interest in things of the mind only. The other person, writing in the first person, is a professor of psychology who knew Frazier while they were both working for their doctor's degrees in graduate school.

The student should try to discover what the chief principle for influencing behavior applied here is from the many illustrations of it. And he should not avoid the question, Are either the means or the ends of Walden Two at all desirable?

* Reprinted with the permission of the author and The Macmillan Company from *Walden Two* by B. F. Skinner. Copyright 1948 by B. F. Skinner.

Each of us," Frazier began, "is engaged in a pitched battle with the rest of mankind."

"A curious premise for a Utopia," said Castle. "Even a pessimist like myself takes a more hopeful view than that."

"You do, you do," said Frazier. "But let's be realistic. Each of us has interests which conflict with the interests of everybody else. That's our original sin, and it can't be helped. Now, 'everybody else' we call 'society.' It's a powerful opponent, and it always wins. Oh, here and there an individual prevails for a while and gets what he wants. Sometimes he storms the culture of a society and changes it slightly to his own advantage. But society wins in the long run, for it has the advantage of numbers and of age. Many prevail against one, and men against a baby. Society attacks early, when the individual is helpless. It enslaves him almost before he has tasted freedom. The 'ologies' will tell you how it's done. Theology calls it building a conscience or developing a spirit of selflessness. Psychology calls it the growth of the super-ego.

"Considering how long society has been at it, you'd expect a better job. But the campaigns have been badly planned and the victory has never been secure. The behavior of the individual has been shaped according to revelations of 'good conduct,' never as the result of experimental study. But why not experiment? The questions are simple enough. What's the best behavior for the individual so far as the group is concerned? And how can the individual be induced to behave in that way? Why not explore these questions in a scientific spirit?

"We could do just that in Walden Two. We had already worked out a code of conduct—subject, of course, to experimental modification. The code would keep things running smoothly if everybody lived up to it. Our job was to see that everybody did. Now, you can't get people to follow a useful code by making them into so many jacks-in-the-box. You can't foresee all future circumstances, and you can't specify adequate future conduct. You don't know what will be required. Instead you have to set up certain behavioral processes which will lead the individual to design his own 'good' conduct when the time comes. We call that sort of thing 'self-con-

trol.' But don't be misled, the control always rests in the last analysis in the hands of society.

"One of our Planners, a young man named Simmons, worked with me. It was the first time in history that the matter was approached in an experimental way. Do you question that statement, Mr. Castle?"

"I'm not sure I know what you are talking about," said Castle.

"Then let me go on. Simmons and I began by studying the great works on morals and ethics—Plato, Aristotle, Confucius, the New Testament, the Puritan divines, Machiavelli, Chesterfield, Freud—there were scores of them. We were looking for any and every method of shaping human behavior by imparting techniques of self-control. Some techniques were obvious enough, for they had marked turning points in human history. 'Love your enemies' is an example—a psychological invention for easing the lot of an oppressed people. The severest trial of oppression is the constant rage which one suffers at the thought of the oppressor. What Jesus discovered was how to avoid these inner devastations. His technique was to *practice the opposite emotion*. If a man can succeed in 'loving his enemies' and 'taking no thought for the morrow,' he will no longer be assailed by hatred of the oppressor or rage at the loss of his freedom or possessions. He may not get his freedom or possessions back, but he's less miserable. It's a difficult lesson. It comes late in our program."

"I thought you were opposed to modifying emotions and instincts until the world was ready for it," said Castle. "According to you, the principle of 'love your enemies' should have been suicidal."

"It would have been suicidal, except for an entirely unforeseen consequence. Jesus must have been quite astonished at the effect of his discovery. We are only just beginning to understand the power of love because we are just beginning to understand the weakness of force and aggression. But the science of behavior is clear about all that now. Recent discoveries in the analysis of punishment—but I am falling into one digression after another. Let me save my explanation of why the Christian virtues—and I mean merely the Christian techniques of self-control—have not disappeared from the face of the earth, with due recognition of the fact that they suffered a narrow squeak within recent memory.

✓
"When Simmons and I had collected our techniques of control, we had to discover how to teach them. That was more difficult. Current educational practices were of little value, and religious practices scarcely any better. Promising paradise or threatening hell-fire is, we assumed, generally admitted to be unproductive. It is based upon a fundamental fraud which, when discovered, turns the individual against society and nourishes the very thing it tries to stamp out. What Jesus offered in return for loving one's enemies was heaven *on earth*, better known as peace of mind.

"We found a few suggestions worth following in the practices of the clinical psychologist. We undertook to build a tolerance for annoying experiences. The sunshine of midday is extremely painful if you come from a dark room, but take it in easy stages and you can avoid pain altogether. The analogy can be misleading, but in much the same way it's possible to build a tolerance to painful or distasteful stimuli, or to frustration, or to situations which arouse fear, anger or rage. Society and nature throw these annoyances at the individual with no regard for the development of tolerances. Some achieve tolerances, most fail. Where would the science of immunization be if it followed a schedule of accidental dosages?

"Take the principle of 'Get thee behind me, Satan,' for example," Frazier continued. "It's a special case of self-control by altering the environment. Subclass A 3, I believe. We give each child a lollipop which has been dipped in powdered sugar so that a single touch of the tongue can be detected. We tell him he may eat the lollipop later in the day, provided it hasn't already been licked. Since the child is only three or four, it is a fairly diff—"

"Three or four!" Castle exclaimed.

"All our ethical training is completed by the age of six," said Frazier quietly. "A simple principle like putting temptation out of sight would be acquired before four. But at such an early age the problem of not licking the lollipop isn't easy. Now, what would you do, Mr. Castle, in a similar situation?" ✓

"Put the lollipop out of sight as quickly as possible." ✓

"Exactly. I can see you've been well trained. Or perhaps you discovered the principle for yourself. We're in favor of original inquiry wherever possible, but in this case we have a more important goal and we don't hesitate to give verbal help. First of all, the chil-

dren are urged to examine their own behavior while looking at the lollipops. This helps them to recognize the need for self-control. Then the lollipops are concealed, and the children are asked to notice any gain in happiness or any reduction in tension. Then a strong distraction is arranged—say, an interesting game. Later the children are reminded of the candy and encouraged to examine their reaction. The value of the distraction is generally obvious. Well, need I go on? When the experiment is repeated a day or so later, the children all run with the lollipops to their lockers and do exactly what Mr. Castle would do—a sufficient indication of the success of our training.”

“I wish to report an objective observation of my reaction to your story,” said Castle, controlling his voice with great precision. “I find myself revolted by this display of sadistic tyranny.”

“I don’t wish to deny you the exercise of an emotion which you seem to find enjoyable,” said Frazier. “So let me go on. Concealing a tempting but forbidden object is a crude solution. For one thing, it’s not always feasible. We want a sort of psychological concealment—covering up the candy by paying no attention. In a later experiment the children wear their lollipops like crucifixes for a few hours.”

“‘Instead of the cross, the lollipop,
About my neck was hung,’ ”

said Castle.

“I wish somebody had taught me that, though,” said Rodge, with a glance at Barbara.

“Don’t we all?” said Frazier. “Some of us learn control, more or less by accident. The rest of us go all our lives not even understanding how it is possible, and blaming our failure on being born the wrong way.”

“How do you build up a tolerance to an annoying situation?” I said.

“Oh, for example, by having the children ‘take’ a more and more painful shock, or drink cocoa with less and less sugar in it until a bitter concoction can be savored without a bitter face.”

“But jealousy or envy—you can’t administer them in graded doses,” I said.

✓ "And why not? Remember, we control the social environment, too, at this age. That's why we get our ethical training in early. Take this case. A group of children arrive home after a long walk tired and hungry. They're expecting supper; they find, instead, that it's time for a lesson in self-control: they must stand for five minutes in front of steaming bowls of soup.

"The assignment is accepted like a problem in arithmetic. Any groaning or complaining is a wrong answer. Instead, the children begin at once to work upon themselves to avoid any unhappiness during the delay. One of them may make a joke of it. We encourage a sense of humor as a good way of not taking an annoyance seriously. The joke won't be much, according to adult standards—perhaps the child will simply pretend to empty the bowl of soup into his upturned mouth. Another may start a song with many verses. The rest join in at once, for they've learned that it's a good way to make time pass."

Frazier glanced uneasily at Castle, who was not to be appeased.

"That also strikes you as a form of torture, Mr. Castle?" he asked.

"I'd rather be put on the rack," said Castle.

"Then you have by no means had the thorough training I supposed. You can't imagine how lightly the children take such an experience. It's a rather severe biological frustration, for the children are tired and hungry and they must stand and look at food; but it's passed off as lightly as a five-minute delay at curtain time. We regard it as a fairly elementary test. Much more difficult problems follow."

"I suspected as much," muttered Castle.

"In a later stage we forbid all social devices. No songs, no jokes—merely silence. Each child is forced back upon his own resources—a very important step."

"I should think so," I said. "And how do you know it's successful? You might produce a lot of silently resentful children. It's certainly a dangerous stage."

"It is, and we follow each child carefully. If he hasn't picked up the necessary techniques, we start back a little. A still more advanced stage"—Frazier glanced again at Castle, who stirred uneasily—"brings me to my point. When it's time to sit down to the soup, the children count off—heads and tails. Then a coin is tossed and if it comes up

SOBAY W B LIBRARY

Box

Access No.



heads, the 'heads' sit down and eat. The 'tails' remain standing for another five minutes."

Castle groaned.

"And you call that envy?" I said.

"Perhaps not exactly," said Frazier. "At least there's seldom any aggression against the lucky ones. The emotion, if any, is directed against Lady Luck herself, against the toss of the coin. That, in itself, is a lesson worth learning, for it's the only direction in which emotion has a surviving chance to be useful. And resentment toward things in general, while perhaps just as silly as personal aggression, is more easily controlled. Its expression is not socially objectionable."

Frazier looked nervously from one of us to the other. He seemed to be trying to discover whether we shared Castle's prejudice. I began to realize, also, that he had not really wanted to tell this story. He was vulnerable. He was treading on sanctified ground, and I was pretty sure he had not established the value of most of these practices in an experimental fashion. He could scarcely have done so in the short space of ten years. He was working on faith, and it bothered him.

I tried to bolster his confidence by reminding him that he had a professional colleague among his listeners. "May you not inadvertently teach your children some of the very emotions you're trying to eliminate?" I said. "What's the effect, for example, of finding the anticipation of a warm supper suddenly thwarted? Doesn't that eventually lead to feelings of uncertainty, or even anxiety?"

"It might. We had to discover how often our lessons could be safely administered. But all our schedules are worked out experimentally. We watch for undesired consequences just as any scientist watches for disrupting factors in his experiments.

"After all, it's a simple and sensible program," he went on in a tone of appeasement. "We set up a system of gradually increasing annoyances and frustrations against a background of complete serenity. An easy environment is made more and more difficult as the children acquire the capacity to adjust."

"But *why*?" said Castle. "Why these deliberate unpleasantnesses—to put it mildly? I must say I think you and your friend Simmons are really very subtle sadists."

"You've reversed your position, Mr. Castle," said Frazier in a sudden flash of anger with which I rather sympathized. Castle was calling names, and he was also being unaccountably and perhaps intentionally obtuse. "A while ago you accused me of breeding a race of softies," Frazier continued. "Now you object to toughening them up. But what you don't understand is that these potentially unhappy situations are never very annoying. Our schedules make sure of that. You wouldn't understand, however, because you're not so far advanced as our children."

Castle grew black.

"But what do your children get out of it?" he insisted, apparently trying to press some vague advantage in Frazier's anger.

"What do they get out of it!" exclaimed Frazier, his eyes flashing with a sort of helpless contempt. His lips curled and he dropped his head to look at his fingers, which were crushing a few blades of grass.

"They must get happiness and freedom and strength," I said, putting myself in a ridiculous position in attempting to make peace.

"They don't sound happy or free to me, standing in front of bowls of Forbidden Soup," said Castle, answering me parenthetically while continuing to stare at Frazier.

"If I must spell it out," Frazier began with a deep sigh, "what they get is escape from the petty emotions which eat the heart out of the unprepared. They get the satisfaction of pleasant and profitable social relations on a scale almost undreamed of in the world at large. They get immeasurably increased efficiency, because they can stick to a job without suffering the aches and pains which soon beset most of us. They get new horizons, for they are spared the emotions characteristic of frustration and failure. They get—" His eyes searched the branches of the trees. "Is that enough?" he said at last.

"And the community must gain their loyalty," I said, "when they discover the fears and jealousies and diffidences in the world at large."

"I'm glad you put it that way," said Frazier. "You might have said that they must feel superior to the miserable products of our public schools. But we're at pains to keep any feeling of superiority or contempt under control, too. Having suffered most acutely from it myself, I put the subject first on our agenda. We carefully avoid any joy in a personal triumph which means the personal failure of some-

U.S. NATIONAL LIBRARY

DATE

ACCE. NO.



body else. We take no pleasure in the sophistical, the disputative, the dialectical." He threw a vicious glance at Castle. "We don't use the motive of domination, because we are always thinking of the whole group. We could motivate a few geniuses that way—it was certainly my own motivation—but we'd sacrifice some of the happiness of everyone else. Triumph over nature and over oneself, yes. But over others, never."

"You've taken the mainspring out of the watch," said Castle flatly.

"That's an experimental question, Mr. Castle, and you have the wrong answer."

Frazier was making no effort to conceal his feeling. If he had been riding Castle, he was now using his spurs. Perhaps he sensed that the rest of us had come round and that he could change his tactics with a single holdout. But it was more than strategy, it was genuine feeling. Castle's undeviating skepticism was a growing frustration.

"Are your techniques really so very new?" I said hurriedly. "What about the primitive practice of submitting a boy to various tortures before granting him a place among adults? What about the disciplinary techniques of Puritanism? Or of the modern school, for that matter?"

"In one sense you're right," said Frazier. "And I think you've nicely answered Mr. Castle's tender concern for our little ones. The unhappinesses we deliberately impose are far milder than the normal unhappinesses from which we offer protection. Even at the height of our ethical training, the unhappiness is ridiculously trivial—to the well-trained child.

"But there's a world of difference in the way we use these annoyances," he continued. "For one thing, we don't punish. We never administer an unpleasantness in the hope of repressing or eliminating undesirable behavior. But there's another difference. In most cultures the child meets up with annoyances and reverses of uncontrolled magnitude. Some are imposed in the name of discipline by persons in authority. Some, like hazings, are condoned though not authorized. Others are merely accidental. No one cares to, or is able to, prevent them.

"We all know what happens. A few hardy children emerge, particularly those who have got their unhappiness in doses that could be swallowed. They become brave men. Others become sadists or

masochists of varying degrees of pathology. Not having conquered a painful environment, they become preoccupied with pain and make a devious art of it. Others submit—and hope to inherit the earth. The rest—the cravens, the cowards—live in fear for the rest of their lives. And that's only a single field—the reaction to pain. I could cite a dozen parallel cases. The optimist and the pessimist, the contented and the disgruntled, the loved and the unloved, the ambitious and the discouraged—these are only the extreme products of a miserable system.

"Traditional practices are admittedly better than nothing," Frazier went on. "Spartan or Puritan—no one can question the occasional happy result. But the whole system rests upon the wasteful principle of selection. The English public school of the nineteenth century produced brave men—by setting up almost insurmountable barriers and making the most of the few who came over. But selection isn't education. Its crops of brave men will always be small, and the waste enormous. Like all primitive principles, selection serves in place of education only through a profligate use of material. Multiply extravagantly and select with rigor. It's the philosophy of the 'big litter' as an alternative to good child hygiene. R

"In Walden Two we have a different objective. We make every man a brave man. They all come over the barriers. Some require more preparation than others, but they all come over. The traditional use of adversity is to select the strong. We control adversity to build strength. And we do it deliberately, no matter how sadistic Mr. Castle may think us, in order to prepare for adversities which are beyond control. Our children eventually experience the 'heart-ache and the thousand natural shocks that flesh is heir to.' It would be the cruelest possible practice to protect them as long as possible, especially when we *could* protect them so well."

Frazier held out his hands in an exaggerated gesture of appeal.

"What alternative *had* we?" he said, as if he were in pain. "What else could we do? For four or five years we could provide a life in which no important need would go unsatisfied, a life practically free of anxiety or frustration or annoyance. What would you do? Would you let the child enjoy this paradise with no thought for the future—like an idolatrous and pampering mother? Or would you relax control of the environment and let the child meet accidental

H. B. F. W. B. LIBRARY

Date

Access No.



frustrations? *But what is the virtue of accident?* No, there was only one course open to us. We had to *design* a series of adversities, so that the child would develop the greatest possible self-control. Call it deliberate, if you like, and accuse us of sadism; there was no other course." Frazier turned to Castle, but he was scarcely challenging him. He seemed to be waiting, anxiously, for his capitulation. But Castle merely shifted his ground.

"I find it difficult to classify these practices," he said. Frazier emitted a disgruntled "Ha!" and sat back. "Your system seems to have usurped the place as well as the techniques of religion."

"Of religion and family culture," said Frazier wearily. "But I don't call it usurpation. Ethical training belongs to the community. As for techniques, we took every suggestion we could find without prejudice as to the source. But not on faith. We disregarded all claims of revealed truth and put every principle to an experimental test. And by the way, I've very much misrepresented the whole system if you suppose that any of the practices I've described are fixed. We try out many different techniques. Gradually we work toward the best possible set. And we don't pay much attention to the apparent success of a principle in the course of history. History is honored in Walden Two only as entertainment. It isn't taken seriously as food for thought. Which reminds me, very rudely, of our original plan for the morning. Have you had enough of emotion? Shall we turn to intellect?"

Frazier addressed these questions to Castle in a very friendly way and I was glad to see that Castle responded in kind. It was perfectly clear, however, that neither of them had ever worn a lollipop about the neck or faced a bowl of Forbidden Soup.

ARTHUR W. MELTON
University of Michigan

The Science of Learning and the Technology of Educational Methods*

"Educational engineering," and all that the phrase implies, may seem to violate the convictions of those who object to manipulating and shaping the individual to fit the requirements of society. That this has been, and is, the function of education in all societies, both historians and anthropologists would agree. If the education student should contend that shaping the society to the individual is a basic tenet of our democratic faith, one can reply that social responsibility is another basic tenet of the same democratic faith. And it is difficult to dispute the use of knowledge that we already have, when, in using it, we can solve some very annoying problems and, perhaps, make ours a better society in which to live.

This article should help the student enlarge his understanding of the relationship of education to psychology, especially to experimental psychology. The relationship of education to the psychology of learning, we are told, is similar to that which exists between engineering and the physical sciences. The psychology of learning is concerned with all the behavioral changes which result from experience, from creative problem solving to our most firmly-held political bias to holding a spoon properly. Education is simply a means for society to manage what happens in the classroom in order to bring about those changes in behavior which it deems most desirable. Melton fully endorses the union of psychology and education;

* Reprinted and abridged with the permission of the author and the *Harvard Educational Review* from the article of the same title, Vol. 29 (1959), 96-106.

and he proposes a program of action. Although he sees several obstacles ahead, his proposals still may be a means of getting the construction of *Walden Two* underway.

In response to the question, "Can the laws of learning be applied in the classroom?" I have chosen to discuss the relationship of the science of learning, as a part of the science of psychology, to the technology of educational methods, as a part of the broader subject which is education. This reflects my conviction that the proper question for consideration is not *whether* the science of learning and the management of learning in the classroom can be mutually supporting, but *how* this desirable—even necessary—relationship can be achieved. What follows is, therefore, an attempt to state some of the assumptions and inferences about the science of learning and the technology of education that make this phrasing of the question necessary, some of the difficulties that stand in the way of achieving the desired mutually supporting relationship, and some ways to overcome those difficulties.

Even though only a part of psychology and a part of education are under consideration, these parts are the core problems of both psychology and education, respectively. On the one hand, the problem of learning—its nature and conditions—is so fundamental to the whole of psychological science that there is frequent, and understandable, confusion between the terms "learning theory" and "behavior theory." On the other hand, the technology of educational methods—if one means by this, as I do, all methods of management of the learning processes of others in order to achieve certain prescribed behaviors or behavior capabilities—is the fundamental technical question in education. There are, of course, other points of intimate and important contact between education and psychological science. It is not my intention to consider them here.

As regards education, the present discussion rests on acceptance of an assumption and two corollaries. The assumption is that educators, whatever their ultimate goals and within whatever social or political system, must know how to manage the learning process in order to achieve the acquisition, retention, and readiness for use of

certain knowledges, skills, and cognitive capabilities. As a first corollary of this assumption, it is assumed that educators must know how to manage the inculcation of attitudes and motives that are necessary for the acquisition and effective utilization of such behavior capabilities, since motivation and learning cannot be considered separately, and the motives of man are largely acquired and thus subject to molding in the educative process. A second corollary is that educators must know how to manage the learning of both behavior capabilities and motivations at all stages of man's development from infancy to adulthood and for all levels of talent, whether that talent is innate or previously acquired.

An assumption about the science of learning and two corollaries also need to be made explicit. These prove to be homologous to those about the technology of educational methods. The assumption is that the science of learning—whatever may be its current limitations of fact and theory—encompasses all forms of relatively permanent modifications of behavior resulting from experience, with perhaps the exception of those modifications commonly identified as sensory adaptation and fatigue. This means that the science of learning must organize our knowledge and understanding of the acquisition of attitudes, motives, affective and emotional responses, mental sets, simple and complex discriminative acts, serial verbal and motor acts, motor and perceptual skills, meanings, concepts and abstractions, and various cognitive capabilities that go under such names as ideational problem solving, thinking, reasoning, decision-making, and even creative invention. In short, I cannot think of any kind of behavior or behavior capability that does not properly belong within the scope of the science of learning if it is a kind of behavior that needs to be managed by education.

The two corollaries of this assumption about the science of learning pertain to the necessary inclusion of: (a) the interactions of learning and motivation; and (b) the interactions of learning and individual differences in talent within the scope of that science. The first of these—the interactions of learning and motivation—has long been recognized by “learning” psychologists as a necessary part of the understanding of learning, and is, in fact, one of the principal reasons it is difficult to distinguish “learning theory” from “behavior theory.” But the interactions between learning and individual dif-

ferences in talent have been relatively neglected in both our learning theories and in our systematic empirical research leading to the analysis and control of the learning process.

These assumptions about the science of learning and the technology of education lead to the necessary inference that the two endeavors have the traditional relationship of science and applied science. Thus, education is to psychology and the social sciences as engineering is to the physical sciences and as medical practice—especially preventive medicine—is to the biological sciences. This analogy—especially the analogy between education and engineering—is one that should be self-consciously adopted and exploited by education. This is not to say that I am advocating that education be described as “human engineering,” especially since that term has been pre-empted by psychologists to describe the applications of psychology (and physiology) to the design of man-machine systems, and this is far removed from the function of education. However, the analogy of engineering to education—one purposefully designing and shaping hardware to the needs of society and the other purposefully designing and shaping human software to the needs of society—proves useful in thinking about the technology of educational methods.

A principal advantage of this conception of education is that it serves notice on all concerned that education has its roots firmly planted in the behavioral sciences, and therefore has a great stake in the acceleration of advances in those sciences. A second advantage is that it becomes self-evident that education is not just the straightforward application of the science of learning, any more than engineering is just a straightforward application of the physical sciences. Just as physicists discovered things that made the creation of television possible, but engineers created television, so likewise psychologists have discovered many things about the learning process, but educational technologists must design curricula and teaching machines that exploit that knowledge. A third advantage is that the model based on the relation between the physical sciences and engineering leads one to expect a relationship of mutual dependency between the science of learning and the technology of education, especially at the frontiers of knowledge of both. Thus, new discoveries in science make new achievements in technology feasible, and

recognized sources of deficiency in technology guide the explorations of scientists. The same sort of give-and-take should occur between the science of learning and the technology of educational methods when the powerful value of the mutually supporting roles of the scientist of learning and the manager of learning are once again recognized by psychology and education.

The wisest first step seems to be to look at some of the difficulties that must be faced if an integration of scientific and technological effort is to be achieved. I refer here to difficulties that are intrinsic to the integration that is to be attempted, not to *ad hominem* difficulties such as the failure of experimental psychologists to be interested in the educational applications of their findings or the failure of educational psychologists to be motivated or equipped to understand and interpret for education the most recent discoveries or the more abstruse theories of learning. These personnel problems will, I believe, diminish or disappear when the partnership is properly conceived and supported. In 1946, for example, few psychologists were involved with the military establishment, as few experimental psychologists are involved in education now. Few would have predicted in 1946 that in 1956 some 700 psychologists, or five percent of the membership of the American Psychological Association, would be engaged in research development, or service associated with the National Military Establishment.

First among these intrinsic difficulties is the present status of the science of learning. While there have been impressive advances in psychology, and especially in our understanding of the learning process during the past 20 years, the fact remains that there is no unified science of learning—and this makes application difficult. Although we know a great deal about classical and instrumental conditioned responses, about selective learning and discrimination, about rote verbal learning, about the learning of perceptual and motor skills, and we are rapidly gaining knowledge about ideational problem solving and decision making, these advances in empirical generalizations and in the theoretical integration of these generalizations have resulted only in *islands* of knowledge and understanding within the science of learning. Furthermore, even within these areas or kinds of learning that have been named, one finds it necessary to place constraints upon our generalizations, because only a limited

number of samples of these kinds of learning have been employed in the investigations. Thus, most of our knowledge about conditioning in human subjects is based upon the eye-blink response or the galvanic skin response; most of our knowledge about the capabilities of human subjects to react appropriately (or somewhat so) to the probabilities of stimulus events is based on simple two-choice guessing situations; much of our knowledge about the learning of perceptual-motor skills is based on tracking or positioning tasks; and most of our knowledge about verbal learning is based on rote serial or paired-associate learning of nonsense syllables or single meaningful words. And finally, it must be recognized that a very large proportion of our knowledge about the learning process has been gained in equally specific experimental situations with sub-human subjects.

Our description of what appear to be the limiting characteristics of the present science of learning should not, in the present context, be considered as a criticism of the way the science of learning has been conducted. There are very good reasons why the most productive scientists in the field of learning have chosen very simple tasks as the focal points of their research and theory, and there is admirable modesty in the effort to build theories that may link two or three of the islands of knowledge before claiming an integration of all the islands (as was customary when psychology was plagued with the *Psychologies of 1930*). Neither should our description be considered as proof that nothing in the science of learning can be used in the management of learning in school situations. The science of learning *can* provide guidance for the management of learning if one has understanding of the science, understanding of the learning to be managed, and ingenuity. However, except in those instances in which the basic scientist has used a practical human task—such as radio code learning, language learning, or target tracking—as the medium for his experimental and theoretical analysis, and this task is also one the learning of which needs to be managed, one cannot expect a cook-book translation of empirical law into management principle.

This brings us to a second, and related, difficulty that must be faced. I refer to the lack, in behavioral science in general and in human psychology in particular, of what may be called a taxonomy of

tasks. I shall not dwell on this beyond a statement of what I mean by it, and what it means for the integration of psychology and education, because it is a topic far beyond the scope of this paper—and furthermore, because I can see the problem but I cannot see the solution for it! My statement means that psychology does not have a satisfactory classification scheme in terms of which specific tasks engaged in by human beings can be described, identified, and placed in a dimensional matrix in relation to other tasks. Without this taxonomy we are forced to use such crude descriptive categories as we referred to previously—discrimination learning, selective learning, tracking, concept formation, paired-associate learning—with the implication that we believe in a typology of learning, when, in fact, most of us do not, and when, in fact, it is known that all instances within these classes are not functionally equivalent.

So, the psychologist, in addition to being plagued with a wealth of variety in human tasks, faces this universe of tasks without even the crutch that would be provided by a systematic taxonomy. This lack of taxonomy places substantial limitations on the ordering of our knowledge about learning and on the feasibility of communicating that knowledge. It therefore stands in the way of the identifications of isomorphisms between learning tasks about which there is information from the laboratories of psychologists and knowledges and skills in which individuals need to be educated. Even tentative refinements in our present crude descriptive classification of tasks would, if combined with an agreement to standardize, greatly improve the communication between, and integration of, the science of learning and each of the education and training technologies that relate to it.

The third major difficulty that must be faced is the prevailing confusion among scientists, technologists, and the users of science and technology regarding basic science and applied science, or technology, on the one hand, and regarding research and operational evaluation, on the other hand. Much has been written and said on this set of problems during the past ten years, largely because the military establishment has been so purposeful in its support of science, and we now have military regulations that define exploratory research, fundamental research, background research, applied research, technical development, and—in the Air Force at least—eight

different phases of testing of the product of technical development. But gross misunderstanding persists. Furthermore, this misunderstanding may be particularly detrimental to the integration of the science and technology in which we are interested, because the science is youthful and the area of application—education—is laden with tradition.

The first misconception that should somehow be avoided is the notion that basic and applied research are antithetical to each other, and that the knowledge gained by basic and applied research does not join together to constitute the science of learning. Such contrasting of the two leads to the implication that both cannot be prosecuted within a single research program, and that different kinds of people with different skills and capacities are engaged in each. These and many other false implications will be avoided if it is remembered that basic and applied research lie on a single continuum, and that the variable involved is the freedom of the investigator in the manipulation of independent variables. Obviously, the more constraints one places on the freedom of the investigator to manipulate variables, the more restricted will be the information gained from the research. Much basic research that goes on in our laboratories of psychology in universities suffers constraints, such as the non-availability of subjects other than college sophomores or the limitation of subject availability to one hour, which act to limit the scientific process and restrict the generality of the findings; on the other hand it is not difficult to conceive of a truly basic research effort on concept formation which would be conducted entirely in a school setting.

This point leads to identification of a second misconception, which is that the instance of learning or kind of learning selected as the object of investigation somehow determines whether the research is basic or applied. This is not true. The science of learning properly includes the understanding of any instance of learning, whether it occurs in paramecia, rats, monkeys, children, or human adults; whether it occurs in the speed of running, in learning a discrimination, in learning nonsense syllables or English-Russian equivalents by rote, in understanding the meaning of "dog" or "cosine," or in discovering the solution to a problem previously unsolved by man. If this view is accepted, then it should be evident that the scientific analysis of the learning of school subjects by six-

year-old children is not, for reason of its practical value, necessarily applied science, as compared with, say, the learning of the barpressing habit by the rat. In fact, within military psychology there have been some quite "basic" programs of research that were at the same time quite directly concerned with the learning of military tasks—such as the programs on the perceptual skills in target detection, on perceptual-motor skills in tracking, and on ideational skills in malfunction-diagnosis in complex equipment.

My next point relates to the over-valuation of applied research when research is being conducted for a purpose, such as the improvement of educational methods. By applied research, I mean, of course, research that is far over toward the end of the continuum where the investigator works with little freedom in the manipulation of independent variables. This usually reduces to a matter of comparing Method A and Method B in the learning of spelling, comparing Training Aid A and Training Aid B for learning history, or in the military, comparing learning with the assistance of a training device and learning without a training device, etc. There was a time, for example, when military psychologists believed that every training device should be evaluated by such an empirical study. A number of us who have been associated with large research programs with applied goals have become very skeptical of such studies. They are usually very costly, and the information gained from them is relatively small in amount and very difficult to interpret. Even if one has been quite careful to obtain a number of independent criteria in terms of which to evaluate the difference between A and B, the information gained is very small compared to that obtained from more analytic, experimental approaches to the problem. The alternative to such applied experiments is to undertake the development of a training program which is designed component-by-component on the basis of sub-programs of highly analytic research studies or on the basis of tests of the applicability of previously established generalizations to the specific context of application, e.g., the age of the learner, the nature of the material being learned, the type of learning involved, etc. Then the final step in such a research and development program is what might be called an "operational suitability" test in which the whole training program is put into actual schoolroom use and the outcomes are routinely measured against the outcomes of previous operational programs. ✓

The final difficulty that I think must be faced in the attempt to integrate the science of learning and the technology of education is that of gaining access to children of school age for the types of experimental investigations that must be accomplished if the science of learning, as it has developed in our laboratories of psychology, is to move in the direction of greater communality with the technology of education. By and large, the organisms most commonly employed in studies of learning by experimental psychologists have been rats, sub-human primates, and college students. Even the military research program has not helped matters too much, because the subjects of these studies have been generally high-school juniors or older (even though some gain resulted from the fact that the populations used have been much more heterogeneous than college students with respect to mental ability). This concentration of effort on samples of human subjects not representative of those whose learning is to be managed in the schools, is not—I am sure—by the choice of the psychologists, but reflects, instead, the difficulty encountered in gaining ready access to school populations for analytic studies that might or might not have any apparent relevance to the proper functions of the schools.

ROBERT M. GAGNÉ, *Princeton University*

ROBERT C. BOLLES, *University of Pennsylvania*

A Review of Factors in Learning Efficiency*

This article reviews the conditions of learning which experimental psychologists have, with some success, attempted to control. It should serve as a useful outline for the student as he

* Reprinted and abridged with permission from *Automatic Teaching: The State of the Art*, E. H. Galanter, ed., 1959, John Wiley & Sons, Inc., pp. 13-53.

turns to later readings which are concerned with particular conditions, such as motivation and reinforcement. The emphasis of the authors on the practical situation rather than on the study of learning in the laboratory setting is also important. When practical considerations enter the picture, the researcher must consider the questions of "learning efficiency" and "transfer," terms with which education students should become familiar, since these are major concerns for teachers.

The student should take special note of the authors' classification of learning conditions into *readiness factors* and *associative factors*. This classification does some of the work of structuring the science of learning, as Melton has urged (pp. 25-27). Many of the readings in this book deal with *readiness factors*, particularly those of motivation, reinforcement, intelligence, and individual differences, as well as others. Other readings, those on cognitive learning and, to some extent, those on communication and social learning, could be classified under *associative factors*.

By referring to this article the education student can begin to organize his own knowledge about the conditions of learning situations. In a more general way the article may help to develop a concept of teaching in which teaching is more a science than an art.

From the very great amount of research that has been done on human learning much is known about the conditions that influence learning, and many of the variables that govern learning have now been identified. It is somewhat surprising that in spite of this body of information, relatively little of a systematic nature is known about how to promote efficient learning in practical situations.

There are probably several reasons for this discrepancy. First, much of the experimental research has been directed toward testing theoretical points which have little immediate practical application. The researcher typically is concerned with understanding how the learning process functions, and not with the question of how to implement learning.

Second, laboratory studies frequently demonstrate the effect of some variable influencing learning by providing conditions that lead to a decrement in performance. It is not altogether obvious that the conditions that *facilitate* learning can be safely inferred from such studies.

Experimental studies of learning have tended to involve rather restricted stimulus material far removed from the kinds of material that are of importance practically. The learning tasks that have been most intensively studied by psychologists have been of an artificial "laboratory" variety; relatively little is known about learning in real life situations. On the other hand, educators, who do work with practical learning situations, have not done the systematic, controlled type of study that is needed to reveal general principles of learning efficiency.

Finally, the criterion of learning employed in most laboratory studies of learning almost always is confined to performance in the learning situation.

For all of these reasons, there appears to be a gap between what is known about learning in the laboratory and learning in the training-job situation. The purpose of this report is to describe and evaluate the findings which might contribute to bridging this gap. Before treating this material directly, we must first discuss several general problems in order to delimit the scope of this report. These problems are: (1) establishing the most useful criteria of learning efficiency; (2) selecting the kinds of tasks for which principles of learning efficiency are to be sought; and (3) a specification of the "structure" of learning, by which is meant a classification of the factors that influence learning efficiency.

Criteria of Learning Efficiency

In seeking to understand the learning process, the psychologist who studies learning typically confines his investigation to a single situation, that in which the learning occurs. He is not usually concerned with what the effects of training in another situation will be. Thus, it is quite natural for him to use, as a criterion of learning efficiency, the number of trials required to produce some arbitrary standard of performance on the learning task.

But the Air Force is concerned with training a man in one situation to perform on the job in a somewhat different situation. Performance in the initial learning or training situation is of lesser practical importance than the degree of transfer that can be effected to the job situation. The purpose of training is thus to equip the trainee with the ability to perform adequately in a situation that is novel in some respect. Consequently, the present report will primarily emphasize transfer rather than learning per se. In other words, the measure of learning effectiveness we will deal with most frequently is amount of transfer.

There still remains some question regarding the criterion of efficiency. One could refer to a learning or training program as being efficient if it required only a short training period, or if it were inexpensive, or if it were successful with a high proportion of the trainees. While each of these possible criteria has merit and may be of primary consideration for some particular purpose, we shall establish for the purpose of this report the following criterion of learning efficiency: *Learning will be said to be efficient if it leads to a high level of performance in the transfer situation.* In accordance with this view, considerations like amount of training are variables which may have an influence on learning efficiency, but they are not measures of it.

The Manipulable Conditions of the Learning Situation

Efficiency of learning clearly depends upon 1) the individual who does the learning, 2) the nature of the task to be learned, and 3) the conditions under which the particular learning occurs. This report will not deal directly with the first two of these three broad categories, although we may note that work proficiency may be greatly enhanced by properly defining the job, and by assigning the proper man to do it. We shall restrict our discussion to the ways in which the learning or training situation itself can be manipulated to produce maximum transfer to the job situation.

Among the conditions of training situations which influence learning and which are accessible, that is, which are manipulable by those in charge of training programs, we may distinguish two classes.

First, there are the motivational or preparatory conditions that make the trainee ready for learning. We shall call these *readiness factors*. These include factors ranging from the general level of motivation to very specific sets to associate particular responses with particular stimuli. Second, there are a number of stimulus conditions to determine which specific associations are formed, and how strong these associations are relative to competing associations. These we call *associative factors*. Various degrees of importance are assigned by different writers to motivational and associative factors. But the evidence indicates that both are important, and while they interact in virtually all learning situations, the present analysis is considerably clarified by treating them separately.

Readiness Factors. Most theorists seem to agree that in learning to perform some task the individual must actively seek some goal or incentive. The individual must be motivated (that is, he must try) to attain some desirable consequence of his performance. Whether this motivation-goal sequence is a necessary condition for learning itself is a much debated theoretical issue; but there is little doubt regarding the efficacy of motivation in producing overt performance.

Allied with conditions of motivation are the conditions of reinforcement, or, put another way, the conditions which govern how goals are actually attained. The effects of reinforcement or goal attainment are complicated; they serve not only to confirm the subject's preceding behavior, but also to maintain the motivational level. Further complications are introduced by the fact that we are interested here primarily in transfer rather than in original learning. Relatively little is known about the role of motivational variables in transfer.

Another important readiness factor in the learning situation is what the learning subject is doing or trying to do. This factor is generally called the subject's *task set*. In general, it can be said that the learner will do better if he knows what he's supposed to do. Task set may be quite general, involving only knowledge of what the completed task is like, or it may be quite specific, as when the person gets ready to press a key at a given signal.

Associative Factors. A second broad class of variables comprises associative factors. These are the stimulus conditions that enter into the learning situation because they are the ones with which specific

responses are to be associated. According to one somewhat oversimplified picture, the problem of controlling behavior consists simply in strengthening association between some stimulus and the desired response to the point where the response will automatically occur whenever the stimulus is presented. While we recognize the possibility of this simple kind of mediation we believe it is practicable and useful to take advantage of other possible types of mediation. Thus, we want to consider seriously the efficiency of learning situations in which the desired responses are associated with a variety of stimuli, and are also mediated by the verbal or voluntary processes of the trainee. This diversification of mediation would seem to be especially useful in highly conceptual types of tasks, as well as those in which variable rather than fixed behavior is called for. Accordingly, we shall discuss later in the report principles which relate efficiency of learning to the nature of what is learned. It seems likely that, for some kinds of tasks, it is more important that the trainee "understand the general principles" underlying his work than that he know only something specific about any particular piece of work. Such a training program would call for the acquisition of a mediating process probably not best characterized as a stimulus-response association.

It is well known that if the training task is similar to the job situation then transfer to a final task (of the job) will be directly related to the degree of learning that occurs in the training task. However, to the extent that the training task *differs* from the job situation, initial learning or overlearning will reduce the amount of transfer and thus be inefficient. The precise degree of similarity which determines the transition point, that is, which determines how much learning is most effective, is the crucial parameter here, but one about which we know little. In any particular instance it is an empirical question just how much learning will lead to the most effective transfer. The reason for this practical limitation is that no well-accepted method is available which makes possible the independent measurement of task similarity. In fact, as the present stage of our knowledge we sometimes depend upon the amount of transfer as an index of similarity! Nonetheless, it is generally accepted that in any specific application transfer is improved by increasing similarity.

In the discussion which follows we have distinguished two roles

played by both readiness and associative factors. We conceive both to play a part in determining the similarity between training task and job, and both to be involved in determining the extent of initial learning during the training period. Thus, we think of the ideal training schedule as a two-stage affair in which, first, learning of the training task is optimized, and second, transfer is insured by making the training task maximally similar to that of the job situation.

Readiness Factors

MOTIVATION

The place of motivational concepts in behavior theory is ambiguous. Recently some writers have suggested that introducing these concepts into the explanation of behavior contributes little toward its explanation. These writers suggest that behavior may best be accounted for in terms of detailed descriptions of the conditions under which it is controlled. On the other hand, we have inherited through the years a good deal of evidence from social mores, from casual observation, as well as from psychological laboratories, testifying to the efficacy of controlling behavior by means of controlling what we call motivational variables. Thus, if we seek to produce some particular kind of behavior in a person, we should see to it that the person *wants* to behave in that way.

Such a broad motivational rule would seem to be trivially obvious and beyond question. Probably there is nothing wrong with it as a general principle. Its fault lies in being too nonspecific and too general; it tells us nothing about how to proceed in any given instance. Theoretical psychologists as well as those interested in applied problems are concerned with controlling and predicting behavior under specific circumstances. Hence, it is necessary to abandon the general rule and to seek in any given situation those particular conditions that maximize performance. To the extent that it is possible to describe kinds of nonassociative conditions that contribute to performance, motivational concepts are useful.

We turn now to the question of how to regulate motivational variables in a training situation so as to maximize performance in that situation. On this question there is a good deal of relevant evi-

dence. The classical finding is that performance improves as motivation increases. Thus the most effective training program would be expected to be one for which motivation is maximal. This is probably true provided certain conditions are met. Spence and his co-workers have recently emphasized that before motivation can facilitate performance it is necessary that the correct or otherwise desired behavior be dominant over other possible behavior patterns. Thus, if the trainee's strongest or most probable response is not the desired one, increased motivation will lead to interference and to performance decrement. This follows from Spence's assumption that the effect of motivation is to facilitate indiscriminately any and all behavior which may be going on. Thus, for motivation to lead to superior performance, the responses which are required must be the ones which are dominant in a situation. If this view is correct (and it is still open to some question) it would suggest that the most efficient learning procedure would be one in which the level of motivation increases in the course of training so as to parallel the probability of the desired behavior. The efficacy of such a procedure has not yet been tested experimentally, however.

Intrinsic vs. Extrinsic Motivation. Another proviso to the general rule that motivation facilitates performance is that the motivation should be, in some sense, relevant to the task. There are a number of task goals for which humans can be motivated. The task itself often provides some intrinsic motivation; the material to be learned may be interesting in itself. Task completion often serves as a goal; other things being equal, people desire to complete tasks they have started. The value of task completion is further enhanced if the task is one in which the trainee is ego-involved, so that pride in success at the task becomes a goal. Desire to succeed appears, in fact, to be a highly dependable source of motivation for the learning situation.

By contrast with these sorts of goals there are extrinsic goals. Success in the task at hand may serve as a goal if the trainee is motivated to excel his fellows, to compete with them. Another kind of extrinsic motivation which may be applicable in some situations is the desire to please one's superiors. Still another is the fondness for gambling. The tendency of people to like to gamble is as yet a relatively unexploited possibility in the design of teaching machines.

Some other kinds of extrinsic motivation approach irrelevance. For example, the learner may be motivated by anxiety over possible failure, or over his inability to do as well as his fellow trainees.

During most of the course of learning it is probably important that motivation be relevant, and preferably that it be intrinsic. Once learning has proceeded to a certain level of proficiency, so that the desired behavior is dominant, it may be that the nature of motivation makes little difference; any source of motivation may sustain performance. In any case, the idea that motivation should be intrinsic rests not so much upon the role motivation plays in learning or in performance during learning; rather, it reflects a concern with the transfer criterion. It seems reasonable to suppose that motives and goals intrinsic to the task are more likely to transfer to the job situation. One reason why training performance is frequently an unreliable indication of subsequent job proficiency may be that the trainee's motives so often change between the training and the job situations.

It should be emphasized that most of this discussion is necessarily speculative. To our knowledge, nothing has been done experimentally to demonstrate that motivation during the training has *anything* to do with the degree of transfer to subsequent on-the-job performance. Furthermore, it seems likely that, even if a systematic experimental research program should indicate the nature of the relationship between motivation and transfer, it would still be necessary to determine empirically what practical measures are required to maximize transfer in any given application. Motivational variables are perhaps the most elusive concepts with which psychologists work.

Levels of Aspiration. Related to the factor of motivation to succeed is the concept of "level of aspiration." The difference between the performance an individual *thinks* he can do, and what he actually accomplishes, has been found to be an important motivational variable. Thus, it turns out, that if the person's goal is set too high he may become disappointed at his relative failure to improve subsequently. On the other hand if his goal is set too low that his learnings will not proceed, he will not improve in the task. It is clear that there is an optimum difference between the goal and the trainee's level of aspiration. While this parameter is undoubtedly an im-

portant one in learning, or at least in performance during learning, its relationship to subsequent performance in a transfer situation remains an unexplored problem.

REINFORCEMENT

The problem of motivation and the problem of reinforcement are highly interrelated. Generally, when we know what a person's motive is, we also know how we can reinforce his behavior. When motivation is intrinsic, that is, when it depends in some way upon the nature of the task, relevant reinforcement is provided by giving the learner "knowledge of results." This is a type of motivation-reinforcement sequence that has been studied experimentally, and several conclusions seem pretty clear. One is that reinforcement should be positive rather than negative, constructive rather than destructive. Reinforcement should be immediate. If it is delayed, the trainee's motivation may lag, and also, the reinforcement fails to provide information which he may need in order to learn anything.

As the effects of delay imply, reinforcement appears to serve two functions. One is to sustain motivation, and the other is to provide information, or feedback. According to some writers this feedback or information value of reinforcement is the only function which the consequences of behavior serve. Other theorists ever since Thorndike have contended that the function of reinforcement is in some way to "stamp in" in some literal sense the stimulus-response association. According to this latter view, reinforcement has little or no effect as far as information is concerned. Whatever the truth of this matter may be, it appears practically reasonable and profitable to administer reinforcement as though its informational value were important. There is ample evidence to show that a trainee's performance may be improved if his scores are reported to him, or if his performance is described and he is encouraged to make improvement. It has been found, in fact, that this is one of the most effective ways in which behavior of the trainee can be modified.

A great deal of attention has been paid in recent years, notably by Skinner and his students, to the fact that performance is apt to be facilitated if reinforcement is made probabilistic. It has been found

that under some conditions subjects will work harder if they are reinforced only once in a while rather than upon every occurrence of the desired behavior. It appears, however, that this phenomenon occurs only with respect to performance, and that it is not reflected in superior learning under partial or intermittent reinforcement. It seems doubtful that transfer to a new situation would be improved by this kind of reinforcement schedule. At present, there seems to be no contrary evidence to the general conclusion that learning is facilitated by frequent, immediate, and positive reinforcement.

SET

The factors of set and attention are frequently mentioned by writers on the subject of human learning. In fact, even the layman would be disinclined to quarrel with such statements as "the learner must be set for learning," or "the learner must pay attention." Yet the scientific literature concerning the effects of these factors on learning is not at all voluminous.

Hebb considers set to be a central neural mechanism comparable to a holding circuit. It is a persisting activity which is set up within the central system, and which has its motor effect only when a second sensory input occurs, with which it acts to produce a response. As a simple example, if we say to an individual "Add these numbers," and then provide him with various sets of numbers, each of the responses will represent his attempt at adding (not subtracting, multiplying, or something else).

The factor of *task set* is apparently the same as Thorndike's factor of belonging, by which he meant the learner's knowledge of "what goes with what." Thorndike performed a series of experiments having the following general pattern. First, he instructed subjects to listen to series of orally-presented materials, such as pairs like "afford 21; equip 34" (in a long list), paying attention as they would if listening to a lecture. After running through the list, in which specific pairs occurred with different frequencies, he then asked the subjects to write the answers to questions like "What number came after 'afford'?" and in contrast, to questions such as "What word came after 21?" Evidences of considerable learning were obtained in answers to the first type of question, in which the pairs seem to "be-

long" together. But almost no learning was found to have occurred *between* the pairs, because, he argued, there was no "belonging." Thorndike points out that this evidence demonstrates the inadequacy of sheer contiguity or sheer repetition for learning. Belonging, or what we call a *task set*, must be present.

In practical training situations, inexperienced teachers may unknowingly violate this principle of *task set*. In the training of complex tasks, the teacher may state the ultimate goal of learning (for example, to learn to operate a control panel), and then proceed to "set the student free" to practice. But the principle of *task set* applies to the *individual items* to be learned, and in such circumstances the student may have to engage in a great deal of needless trial and error behavior before discovering for himself "what goes with what" (for example, that a particular knob controls the activity of a particular dial). The importance of this principle is, therefore, that to maximize learning efficiency means must be found, usually by instructions, to establish suitable *task sets* to each of the items of the total task to be learned. Attending to the stimuli which are relevant, as defined by the final task to be performed, is an important condition for the assurance of a high degree of transfer.

Associative Factors

When we turn to a consideration of associative factors in learning efficiency, we need to consider the variables affecting the nature of what is to be learned. More specifically, these are the number, order, and nature of associative connections that can be manipulated within the learning situation. As stated previously, we are here considering learning in terms of performance of a final task which is not necessarily the same as the "materials to be learned." This means that we must deal with variables that have been investigated in connection with *transfer of training*, and not simply learning itself.

In all, we shall consider here three classes of factors in relation to learning efficiency. The first is what is to be associated, or the nature of the associations to be established, considered in relation to the task on which performance is desired. The second is *intra-trial factors*, or those conditions which may be varied systematically within each trial of learning, applying equally to all trials. The

third is *inter-trial factors*, which may be manipulated in some orderly way between learning trials, or in stages as learning proceeds.

THE NATURE OF ASSOCIATIONS

The most important characteristics of associations to be learned, if we keep in mind the transfer of learning to a criterion task, pertain to *similarity*.

Stimulus Similarity. Concerning the factor of stimulus similarity there has never been any serious disagreement of experimental evidence with the following rule: Positive transfer increases with the degree of similarity of the stimuli of the initially learned task to the final task. Thus the significance of this principle for learning efficiency is clear; stimuli of the associations to be learned should be made as nearly like the stimuli of the final task as possible. In terms of practical training situations, this principle may well be tempered by feasibility. For example, if an operator must learn to identify the switches, knobs, and dials on a panel, can these be represented in photographs (or even drawings) rather than as three-dimensional objects? The answer appears to be that high amounts of positive transfer may be obtained by representations of stimulus objects. This means that the stimuli to be associated in the learning situation can be pictured, rather than "real," without great losses in transfer. On the other hand, providing simply conceptual representation for stimuli, as is done when words are used rather than pictures, is another matter entirely; and in such instances the principle of stimulus "similarity" may in fact be violated. A picture of an amplifier may be highly similar to the amplifier itself, but the word "amplifier" as a stimulus is by no means similar. In any case, the principle of stimulus similarity is not one on which the books can or should be closed; it will require a great deal more research to provide precise meaning to this phrase.

Response Similarity. It would be convenient indeed if we could state that there is a comparable rule about the response members of the associations established by learning, namely, that transfer of training increases with the degree of similarity of responses of the initially learned task and the final task. We cannot agree that this

principle should be considered well established. There are several reasons for this:

1. Generally speaking, it is known that the mediating responses for motor acts, acquired in initial learning, need not be highly similar to these motor acts in order for high degrees of transfer to occur. If an individual is able to identify the location of objects in a picture by pointing to them, we expect that he can also walk to them correctly when he is at the actual scene of the picture. Yet the responses in this initial and final task are really quite different. There are not many experiments in this field, probably because the facts have appeared so obvious.

2. Most of the evidence on response similarity has been obtained with the learning of paired associates, and some of the crucial evidence comes from studies employing responses which are similar in meaning. The difficulty may be, that the second member (often called the "response member") of a pair of associates has a stimulus function, as well as being employed as a response. As a consequence, it interacts with other members which are similar in meaning, and it thus has an effect on transfer. But this meaningful similarity is better having as a *mediating stimulus*, rather than as a response pure and simple. How else, in fact, could *meaningful similarity* of responses be interpreted?

3. The results of paired associate learning are strongly influenced by intra-list interference, as has been demonstrated by many studies. It seems particularly doubtful that one can draw valid conclusions about first-task-second-task similarity unless intra-task similarity has been measured separately or ruled out completely. It is probable, therefore, that the empirical results obtained on response similarity in paired associate learning are quite inadequate for significant conclusions to be drawn concerning the effects of this factor on single associations in learning.

Response similarity has been demonstrated as a factor affecting response strength in studies of generalization, as well as in studies of motor skills. But this evidence is an inadequate basis for the derivation of principles of learning of verbal and conceptual tasks. In view of this, and the objections raised to existing evidence as listed above, we must conclude that *response similarity* is a factor concerning which we know very little. The question of learning efficiency

requires a good deal more systematic knowledge regarding the effects of this variable.

Similarity in Serial Tasks. Some special mention needs to be made of the effects of similarity of associated items in sequentially-learned materials, which are relevant particularly to job tasks of following procedures. In such tasks, each element clearly functions as a response and also as a stimulus to be associated with the next succeeding response in the series. As a number of investigations have shown, the learning of sequential verbal material is influenced both by intra-task similarities and by the similarities of the learning task to the final task. In fact, the separation of these two effects has not yet been satisfactorily determined.

The learning of sequential verbal tasks increases in rate as the individual members are made less similar to each other. This suggests that when we have a procedural task characterized by associative interference, the mediating task provided for learning may be facilitated by making the associated members less similar to each other than those of the procedural task itself. In other words, it seems possible that the members of the learning task may be made *more distinctive* than the members of the final task, and thus increase learning efficiency. However, it should be remembered that although the effects of such intra-task variation are predictable, we do not know what effects this treatment would have on inter-task interference; in other words, we do not know its effects on transfer to a job task. It is apparent that, so far as following procedures are concerned, the interplay of similarities among elements and between initial and final tasks is an area in which considerable additional research is needed.

INTRA-TRIAL FACTORS

In continuing our consideration of associative factors, we next turn to a set of variables that may be manipulated within each and every trial of learning. As we have pointed out previously, these are to be distinguished from variables which are systematically varied from trial to trial, or in stages as learning progresses. There are three primary ways in which such intra-trial factors may be manipulated, and we shall discuss them here. For any given response, the stimuli to be associated with it may be varied in number and variety.

Second, for any particular stimulus, the individual may be required to learn different numbers of responses. And third, the meaningfulness of the associations may be varied.

Number of Stimuli. We can adduce little evidence from the experimental literature concerning this factor and its influence on transfer of training. The experimental question may be described as follows. Suppose we are interested in the performance of an identification task of fifteen components (such as, the components of a newly-developed weapon). In a standard learning situation we would expose a picture of each item and require the learner to respond with its name. However, being aware of the effects of similarity in producing associative interference, and thus decreasing the rate of learning, we might decide to add additional stimuli to each pictured item, in order to make them more distinctive from each other. Increased distinctiveness, for example, might be added by accompanying each item with a distinctive color, a distinctive symbol, a distinctive border, etc. It is sometimes argued that the effect of this added number and variety of stimuli would be to reduce associative interference and thus speed up the learning. The question as to whether transfer to the final task would be as good or better under these conditions needs to be answered by experimental investigation. The idea of providing extra "stimulus support" for learning is one of the hypotheses that appears to be involved in the work of Skinner on teaching machines.

Number of Responses. This is another factor concerning which no direct experimental evidence exists. Using the fifteen-item identification task as an example again, the standard learning situation in which each stimulus is associated with a single response (e.g., a name) may be contrasted with one in which additional responses are also required to be learned to the same stimulus. For example, we might require the learner to acquire the responses "square," "black," "voltage," as well as "amplifier" to a picture of an amplifier. Presumably, it would take him longer to learn four responses to each stimulus item than it would for him to learn one. Nevertheless, the experimental question of importance to learning efficiency concerns the matter of the effects of this variable on transfer of training. If the criterion of transfer is employed, it is entirely conceivable that

the added effort (and time) required for initial learning might be overbalanced by advantages in transfer of training.

Also relevant to this question may be the results on meaningfulness, to be discussed below. There can be little doubt about the faster learning of meaningful materials, although an advantage in recall does not appear. Noble hypothesizes a direct relationship between meaningfulness and number of associated responses. If one accepts this notion, then the superior efficiency of meaningful learning materials may be attributed to the greater number of previously acquired associations such stimuli have. Another implication is that the meaningfulness of stimuli (and thus their transfer effectiveness) may be manipulated in the learning situation by requiring the learner to acquire a number of responses to the stimulus. Skinner's technique of "ringing the changes" on a particular principle to be learned may also be basically a matter of increasing the number of responses to single stimuli.

Notwithstanding the existence of this rather indirect experimental evidence, there remains a need for research directly aimed at finding an answer to this question about number of responses and transfer. The possibility exists that requiring the learning of increased numbers of responses to the same stimuli may be a significant factor in learning efficiency, insofar as it can increase positive transfer.

Meaningfulness. As is well known, many investigations of human learning have been concerned with nonsense materials. Frequently, research is conducted with materials for which meaningfulness is held as a constant, preferably low, value. But there is considerable evidence that meaningful materials are learned more rapidly than are nonsense materials. In fact the differences in learning usually found in favor of meaningful materials imply that this is a factor of outstanding importance to learning efficiency. The question of learning efficiency, which has not been directly investigated, may be stated as follows: (1) Given a set of inherently meaningless identifications to be made in a job task, can transfer be most effectively mediated by the acquisition of meaningful associations? (2) Given a more or less meaningless sequence of acts to be performed in following a procedure, can transfer be insured by acquiring a meaningful verbal sequence representing these acts? (3) What effect does degree of meaningfulness of concepts acquired in a learning situa-

tion have upon the performance of "concept-using" tasks including problem solving?

The experimental evidence shows, first of all, that there is a regular increase in rapidity of learning as the material to be learned increases in meaningfulness. This effect is enhanced when the members being associated are connected by some logical sequence. There is a definite relationship between this finding and the long-known effectiveness of mnemonic systems. Cofer's findings on the retention of meaningful materials show that learned concepts may be acquired (as "ideas") much more rapidly than can exact verbal passages, and that they continue to function as concepts for a long time after exact verbal sequences have been forgotten. This finding is consistent with the suggestion of certain writers that mediation by means of meaningful materials may be most economical of learning time because a small absolute amount of material must be acquired. Putting all these things together, there is some doubt that meaningfulness has ever been accorded quite the importance it deserves as a factor in learning within the framework of traditional investigations. Further, the transfer value of meaningful materials has only rarely been emphasized.

INTER-TRIAL FACTORS

There are two general kinds of factors which may be varied between trials of learning. One of these is the temporal distribution of trials or practice sessions, and the other is change in the learning task.

Massed or Distributed Practice. The question of whether practice is more efficient if trials are massed or if they are distributed is a classical problem which has been studied extensively. Basically, it is clear that if the inter-trial interval is too long, everything that has been learned on the preceding trial tends to be forgotten and to have to be relearned on the next trial. Even partial forgetting implies that some relearning could be avoided by reducing the interval between trials. On the other hand, closely massed trials are likely to produce fatigue, boredom, and work decrement. In this latter case, however, it is not altogether certain that the decrements so typically displayed actually indicate a retardation of learning. Some studies

suggest rather strongly that the decrements apply only to the performance, not to the learning. When rest intervals and a suitable test are introduced following the period of learning, it is found that massed practice groups have actually acquired more than they have demonstrated during learning. Thus massed learning is more efficient than it would appear. Probably the main effect of massed practice is upon motivation rather than on association. Several studies of serial learning have found that performance is facilitated to a greater extent by increasing the time per item during the presentation, than by introducing rest intervals between repetitions of the list. This finding suggests that the most efficient learning program, in terms of total elapsed time, may be one in which material is presented slowly within each trial, but massed in the sense that one trial promptly follows another. Finally, there are studies which have suggested that when the task to be learned is a very difficult one, inhibitory effects of massing trials may be more than offset by the difficulty of remembering procedural details from one trial to the next.

If we conceive of the intra-trial interval as being important because it relates to the ease with which the learner can proceed in learning, one obvious solution would be to let the trainee set his own intra-trial interval, that is, let him pace himself. Skinner has argued that this is one virtue of his teaching machines. It may well be that the learner is the best possible judge of his own level of inhibition and his own best judge of when he is ready to proceed in learning. With such a self-pacing procedure it is usually found that the subject chooses a relatively long intra-trial interval early in learning, and decreases this interval as learning proceeds. Again we must note that the temporal sequencing of training which leads to the most efficient transfer to the job situation needs to be determined by experimental study.

Task Scheduling. A more complex question concerns how the relevant stimuli should be scheduled. There appear to be two distinct schools of thought on the subject. One of these contends that the underlying principle (or the crucial stimulus element) with which the behavior is to be associated should be emphasized from the first, and should serve as a stable reference point throughout the course of learning. According to such a position, departure from this pro-

cedure can only lead to interference and learning decrement. In practice, however, this procedure has the disadvantage of frustrating the learner if he cannot make the correct responses at the outset. On the other hand, Skinner, among others, has maintained that the desired response should be given every possible stimulus support from the outset of learning. The purpose of this stimulus support is to insure that the correct response will be made. Once made, it can then be reinforced, and the superfluous stimuli gradually removed. The obvious difficulty with this strategy is that it is quite likely that the reinforcement will strengthen the association of the response to the wrong stimulus. Which procedure leads to the greatest efficiency of learning and the greatest transfer to a new situation, is not presently known.

It seems likely that superiority of one or the other procedure may depend upon many factors. If the correct response is a verbal one, if it can be elicited by instruction or by some other means, and if a suitable task set can draw the trainee's attention to the relevant stimulus, then this would seem to suffice. However, it is clear that there are many learning situations in which all these things cannot be done. This is true, for example, when the correct response and the relevant stimulus cannot be verbally communicated to the trainee. We may note in this connection Skinner's method of using extra "stimulus supports" is derived by analogy from his animal studies. While he argues for the efficacy of this method in teaching machines, we do not know that the two procedures have been put to a critical test. One obvious disadvantage of Skinner's procedure is that once the response has been established in the presence of a number of stimuli, it is then necessary to eliminate the irrelevant stimuli from the total stimulus complex, in order that only the relevant one will remain in association with the desired response. This means that learning must proceed as a discrimination problem in which the relevant stimulus must be discriminated from the irrelevant ones, or else the danger is run that the trainee will conclude his training having learned something irrelevant. Thus, there is the possibility that additional training is required, over that which is really necessary for the desired performance, before the trainee is in a position to go into the transfer situation. Those who uphold Skinner's position might well argue at this point that the superfluous

stimulus support given the correct response early in learning can be gradually withdrawn so that the overall learning efficiency is not impaired. This argument fails to recognize, however, that these irrelevant stimuli not only support the correct response but also incorrect and competing responses which may introduce interference. As an example of this problem consider learning of the touch system of typing. The question we are considering may be phrased as follows: Can the touch system of typing be learned better if (1) the names of the keys are present, or (2) if the keys are blank? Most educators seem to agree that while the presence of the letter names on the keys is an aid in early learning, they may actually impede the ultimate level of proficiency desired in touch typing. More research needs to be done before any clearcut answer to this question can be obtained.

Wolffe has pointed out that in any instance of stimulus-response learning, if the contextual stimuli remain constant through the course of learning, they will all tend to become (irrelevantly) associated with the correct response. On the basis of evidence from several studies, he is led to conclude that a desirable condition for efficient learning would involve the use of a variety of contextual stimulus conditions, in which only the relevant stimulus remains invariant. The evidence indicates that learning under these variable conditions is indeed less resistant to extinction than is learning under constant stimulus conditions. Such evidence appears quite inconsistent with what a "stimulus support" principle would lead us to expect.

Summary and Conclusions

On the whole, our conclusion from this evidence must be that there are few principles which can be directly applied to the problem of making learning efficient. The findings concerning the nature of the learning process in human beings are primarily suggestive for this problem, rather than productive of verified practical rules for the control of conditions of efficient learning. This means that the attempt to manipulate learning conditions, whether carried out by a teacher or by the designer of a teaching machine, must employ a

good deal of art and not much science, at the present stage of knowledge.

On the other hand, our review of the factors in efficient learning shows us that there are quite a number of these factors which may, in any given situation, be manipulated to affect learning efficiency. If these could all be systematically controlled by means of a machine, or by an otherwise well-designed learning situation, the possibilities of increasing the efficiency of learning over that which typically results from practical training appear great indeed. A suitably designed machine could, of course, be used to carry out such a program of research. Estimates can be made of the relative importance of these factors we have described to learning efficiency. But to attain the goal of ultimate control over learning, it is even more important to undertake research which will determine how far each of these variables, or combinations of them, can be pushed in making learning efficient. As we have pointed out, such a question is neither asked nor answered by the conventional experimental study of human learning.

References

- COFER, C. N., A comparison of logical and verbatim learning of prose passages of different lengths. *Amer. J. Psychol.*, 1941, 54, 1-20.
- HEBB, D. O., *A Textbook of Psychology*. Philadelphia: W. B. Saunders Company, 1958.
- NOBLE, C. E., An analysis of meaning. *Psychol. Rev.*, 1952, 59, 421-430.
- SKINNER, B. F., *Science and Human Behavior*. New York: The Macmillan Company, 1953.
- SKINNER, B. F., The science of learning and the art of teaching. *Harvard Educ. Rev.*, 1954, 24, 86-97.
- SKINNER, B. F., Teaching machines. *Science*, 1958, 128, 969-977.
- THORNDIKE, E. L., *The Fundamentals of Learning*. New York: Teachers College, Columbia University, 1932.
- WOLFLE, D., Training. In S. S. Stevens ed., *Handbook of Experimental Psychology*. New York: John Wiley & Sons, Inc., 1951, 1267-1286.

FREDERICK J. McDONALD
Stanford University

The Teacher and the Improvement of Educational Practice*

If the beginning education student discovers that conventional educational practices leave something to be desired, he may well ponder questions about how he is to go about improving them. Does improving classroom practice consist merely in trying out new gimmicks and guessing how successful they have been? The author of this article suggests answers for such questions and, in doing so, outlines a systematic and sober program for continued educational reform. Briefly outlined, a teacher is advised (1) to know specifically what his teaching objectives are, (2) to use his general knowledge about learning situations as a basis for hypothesizing—that *if* he has his students do these things, *then* he can expect these other things to happen, and (3) to test the hypothesis instead of merely guessing how well the students succeeded in achieving the specified objectives.

These three steps comprise a flexible model for almost every act of teaching, and the student can constantly refer to it in the succeeding readings. The reference above to “general knowledge about learning situations” is important, because this book of readings contains articles which are either sources or surveys of sources of such knowledge. The article by Gagné and Bolles (pp. 31–51) is the most complete statement in the book of the characteristics of learning situations which have been studied in the laboratory. Also, the readings on motivation, reinforcement, concept learning, social learning, etc., are attempts to find out about the conditions under which we learn

* Reprinted and abridged with permission of the author and publisher from *Educational Psychology* by Frederick J. McDonald, pp. 682–698. © 1959 by Wadsworth Publishing Company, Inc., Belmont, California.

and their relation to each other. In addition, Chapter Ten is devoted to the question of testing our hypotheses and practices to find out how well we are doing the job we intended.

McDonald encourages the education student to behave as an experimental psychologist by adopting not only scientific procedures but also the scientist's critical attitude. Unless the student learns to challenge conventional practices and what often parades as "common sense," he may never see the need to test constantly his assumptions about and practices in teaching.

The Relative Validity of Generalizations about Practice. The assumption that educational practice can be improved is implicit in our conception of the character of empirical knowledge. We assume that knowledge about observable events and knowledge of propositions which describe and explain these events is essentially *probabilistic* in character. Therefore, since any statement about educational practice is a statement about observable events, it can be only *probably* true.

Suppose, for example, that a teacher uses a film on India as part of a unit on "understanding other cultures." He may find that students are interested in the film and acquire considerable information from it and that they ask intelligent questions about the customs of the Indian culture and show an increased sensitivity to cultural differences. For this teacher, this film "works." That is, when he used the film in a learning experience, behavior changes occurred in the directions he had specified as desirable. Has this teacher invented an educational practice which cannot be questioned, cannot be improved upon, or should not be changed?

The Frame of Reference of a Generalization. These questions can be answered intelligently if we view the propositions about the use of this film as being only *probably* true. Why are they only *probably* true? In the first place, because the film was used within a framework of environmental variables. The teacher may not have analyzed these environmental variables, or even identified them. We can suggest what some of these variables might be. Some might be related to the particular sample of pupils who saw the film; although these

children have many characteristics in common with other children, they may systematically differ from other children on the variables relevant to their ability to learn from the film being used. The intelligence levels of the children, their ages, social class backgrounds, and previous learning experiences are some of the variables which may be relevant to the fact that the film "worked."

There are other contextual variables which might have been relevant to the success of the film, such as the content communicated, the manner of content presentation within the film itself, the presence or absence of emotional appeals in the film, the prestige of a communicator in the film (if one is used), or the manner in which the teacher related the film to other activities. We have not specified all of the possibly influential variables, nor is it necessary to do so for our purpose. The point here is that this film was used in a particular set of conditions, and the validity of the proposition about the success of the film is *probably* high if the film is to be shown only within the context of the same conditions.

Decision-Making and the Relative Validity of Generalizations. Certainly no one will argue seriously that we should not make the best decisions that can be made about educational practice on the basis of our present knowledge of it. Our explicit assumption that educational practice should be improved is a statement that we need to make better decisions—decisions with greater probability of being valid. You will notice that we have not adduced horrendous examples of poor educational practices as evidence that educational practice needs to be improved. We did not do so precisely because we are attempting to point out that decisions about the validity of an educational practice can be made only after critical inquiry which takes account of the changes in behavior specified as desired. An educational practice is not "poor" because it did not "work" nor "good" because it did. Until we have specified what we mean by saying that a practice "works," and until we specify under what conditions and for what purposes a practice is likely to be effective, we are not in a position to label any practice as generally "good" or "poor."

The author recalls being told by an older teacher, at the beginning of his first year of teaching, "Don't smile for the first six months." Some readers may be amused by this suggestion, but the author must confess that he accepted it with some seriousness. For

the purposes of our discussion, notice the character of this proposition. It is a statement about what to do in a classroom setting. In context, it is a suggestion about how to act in order to achieve "pupil control." Implicit in this statement is the hypothesis that if a teacher does not smile during the first six months of school he will be more likely to achieve "pupil control." This proposition could be buttressed with deductions from propositions about the role of the teacher in relation to his pupils, the pupils' perceptions of a young teacher, and other generalizations about teacher-pupil relations. Is the suggested educational practice a "good" or a "poor" practice? Assume that the author had followed this suggestion (which he did not) and found that he was able to maintain an unspecified state of affairs called "pupil control." Could we say that he had discovered a practice that guarantees success in "maintaining order"?

Formally, this proposition is a statement about a set of observable events and, in principle, it can be only probably true. We would not contest the fact that a given teacher did not smile for six months nor that he maintained "pupil control." Our proposition is drawn from two specific sets of events, stating a relation between them; as a generalization, however, it is meant to be applied to a wide variety of comparable situations. Furthermore, there may be other generalizations which are more probably valid than this one. Educational practice can be improved if we obtain these more valid generalizations because decisions for practice based on them are more likely to be valid.

Decision-Making and Improved Prediction of Behavior Change

We can summarize the above ideas while introducing a new conception. Educational practices are based on propositions about sets of observable events, propositions which describe or explain the relations between learning experiences and pupil change; they can be used as bases for the decisions required to create learning environments. The use of more valid generalizations in the decision-making process improves educational practice by strengthening our ability to predict and control the events with which we are concerned.

Assume that we find that a more valid statement about the relationship between teacher behavior and pupil behavior to be this: "Pupil interest in classroom activity is related to the degree in which teachers manifest a personal interest in their pupils." This generalization enables us to improve our predictions about the relationship between teacher behavior and pupil behavior. It suggests that if a teacher takes a personal interest in his pupils, the students will probably show increased interest in class activities. It directs us toward more valid information about the variables related to pupil interest in class activities.

We have noted two characteristics of propositions that describe educational practices: (1) they are generalizations about relationships between sets of events; and (2) as generalizations, they are inferences from observations made within specific contexts. The validity of these propositions depends in part upon the context of conditions within which the observations were made. Implicit in any proposition describing educational practice is a set of conditions upon which its validity depends. These sets of conditions specify a frame of reference within which the generalization is said to be probably valid.

Applying Generalizations in Making Specific Decisions. The importance of the conditions on which the validity of a generalization depends is apparent when we analyze the requirements for making decisions about specific educational practices. Assume that the teacher has formed a generalization about the relationship between his behavior and the behavior of his pupils. He now wishes to "apply" this generalization in making decisions about specific practices in his classroom. Before he can "apply" this generalization to his own practice, he must assess the probability that the defining conditions of the generalization are present in his classroom. The generalization was derived from the experiences of a particular sample of pupils (with all their individual characteristics) and a particular teacher (with his own individual characteristics). Furthermore, the generalization probably applies only to certain kinds of teacher and pupil behavior.

You will recall that we analyzed the generalization "Praise improves pupil performance." This generalization was derived from data obtained in an experimental situation; this situation defined

the conditions under which the generalization would probably be true. The teacher will have to assess the extent to which these conditions are present in his classroom before applying this generalization in the context of the learning environments he is creating.

Before making this application, the teacher should formulate a second level of hypotheses. These hypotheses are more specific than the generalization and apply to the context of the particular learning experience. One such hypothesis might be: "If I praise the pupils in my class for correct answers when they are working on arithmetic problems, performance in arithmetic is likely to improve." The validity of this derived hypothesis is unknown. We predict that it is likely to be valid insofar as the conditions in this classroom are similar to the conditions under which the original generalization was derived. Coladarci makes this distinction between levels of hypotheses in the following way:

If we knew for certain what the correct manipulations were [manipulations of the learning environment to produce behavior change], we would have an easy solution to questions about educational procedure—we could have valid "rules" to follow. We frequently do not have such knowledge. Even our most rigorous and competent research provides conclusions that are properly stated only as probabilities.¹

The practitioner must translate these probabilities into a secondary set of probabilities about what would happen in the case of his purposes and his pupils. He must make inferences about the application of these probabilities to his teaching methods, curriculum, organization, and so on. The educator's operations are best thought of as hypotheses and, like any hypotheses, they must be tested.

The teacher's operations, as an educational practitioner, are analogous to those of a medical practitioner. In the course of his training the medical practitioner has learned certain generalizations about practices likely to cure illnesses; but he must also make second-order hypotheses about the applicability of these generalizations to specific patients. For example, he may have learned a generalization about the likelihood of sulfa drugs curing certain kinds of infections. When he is treating a patient with an infection, he does not

¹ A. P. Coladarci, "Research in a Local District," in *Educational Research in Local School Districts*, Research Résumé No. 3, San Francisco, State Advisory Council on Educational Research, California Teachers Association, 1957, p. 28.

act on the generalization until he makes a hypothesis that the use of a certain sulfa drug will cure this particular infection in this individual patient. If it does not, the practitioner need not assume that the generalization was necessarily incorrect, but rather that the conditions under which it is correct were not fulfilled in this particular case. The teacher, like any practitioner, operates within a specific context and with specific people.

The Teacher and the Validation of Hypotheses. This discussion should suggest to the reader that two levels of critical inquiry are necessary for the improvement of educational practice: at one level, we should test generalizations that are applicable to a wide variety of particular events; at the other, we should test each specific hypothesis derived from a comprehensive generalization.

The first of these tasks is properly the work of trained investigators. Comprehensive generalizations can be developed from data gathered according to careful experimental designs; this kind of systematic research requires skills which most classroom teachers do not have. But critical inquiry into the validity of hypotheses derived from these generalizations is a task for the classroom teacher. Thousands of teachers in thousands of classrooms make numerous decisions each day about the effects of learning experiences on pupil behavior. The likelihood is small that each of these hypotheses will be investigated with formal experimental procedures. But if teachers do not assume the responsibility for a critical evaluation of their own procedures, it is not likely that educational practice will be improved—they will remain mystified as to why what “worked” yesterday does not “work” today. Without such critical evaluations, teachers will be unable to support or defend their methods, and hunches, guesses, and recipes of unknown validity will continue to be the sources of decisions about educational practice.

Procedures for Critical Evaluation of Educational Practice

As we noted above, the systematic investigation of educational practice requires a knowledge of formal experimental procedures. While teachers may not be able to evaluate their derived hypotheses in the systematic manner that characterizes formal research, certain

principles of research are relevant to the informal evaluations that the teacher can make of his hypotheses. In the following sections we will outline some basic principles of critical inquiry that may be useful to the teacher in analyzing his own educational practices.

Developing Testable Hypotheses. Propositions about educational practice are, in principle, hypotheses about conditions likely to produce behavior change. If these hypotheses are to be evaluated, they must be stated in a testable form. Recall the generalization that we used as an example of a proposition assumed to be obvious. This proposition is essentially untestable in the form in which it was stated. If we gave tests to determine how much teachers knew about the subjects they were teaching and determined that they had limited knowledge of them, we still would not have tested the hypothesis. We would have obtained information about their knowledge of subject matter, but we would not have determined the effectiveness of their instruction.

A hypothesis relates two variables—in this case, “good teaching” and “knowledge of subject matter.” Until we have measured both of these variables and determined the relationship between them, we have not tested the hypothesis. Before doing this, we must define “good teaching” and “knowledge of subject matter.” Once we have formulated these definitions, it may be possible to develop measurement procedures for discerning degrees of “good teaching” and “knowledge of subject matter.” At this point it is possible to test the relationship between the two variables.

Some hypotheses are untestable because they relate variables which are outside the domain of empirical events. For example, the hypothesis that pupil misbehavior is caused by “devils” is untestable in principle. Only one of the variables in this proposition is measurable—namely, the extent to which pupils misbehave. The other variable, “infestation with devils,” cannot be defined in empirical terms and, consequently, cannot be measured in any way.

A less absurd example, but one representing an equally untestable hypothesis, would be the following: “Inspiration promotes creativity.” We can define “creativity” in empirical terms, difficult as this may be, but “inspiration” is such a vague concept that we cannot tell whether there are any empirical referents for it. Indeed, we probably cannot observe the variable called “inspiration” without

confounding it with the variable "creativity." We cannot test such hypotheses.

The Formulation of Hypotheses. There are two general processes by which hypotheses are developed: induction and deduction. An *inductive* hypothesis is formulated as a generalization from observed relationships. Teachers who have noticed that students are restless during a rainy season might make this inductive hypothesis: "Rainy weather produces restlessness in children." Using the process of induction does not guarantee the validity of the generalization. As we have noted consistently throughout this book, inferences must be checked, and inductive hypotheses are inferences.

The hypothesis about restlessness could also be tested by making deductions from it and then testing these deductions as minor hypotheses. From our original hypothesis we might deduce that students would be less restless during other seasons of the year. If this deductive hypothesis has validity, we infer that the general hypothesis from which it was deduced also has some validity. However, before we can determine the validity of the more general hypothesis, we must develop a whole series of deductive hypotheses, each of which must be tested.

As the preceding example has shown, a deductive hypothesis is drawn from, and is consistent with, some more comprehensive generalization or proposition. Travers has provided an interesting example of deductive hypotheses from more comprehensive statements of relationships (see Table 1).²

Table 1

A THEORY OF THE EARLY STAGES OF LEARNING READING
(FROM TRAVERS)

Definitions

1. *Reading* is defined as a controlled form of talking in which the words that are said are controlled by the nature of the written symbols presented.
2. *A correct reading response* is defined as the act of saying the agreed-upon interpretation of the written symbol presented.
3. *Accuracy of response* to a word is defined as the percentage of attempts to say the words that are correct.

² R. W. Travers, *An Introduction to Educational Research*. New York: The Macmillan Company, 1958. Reprinted with the permission of the publisher.

4. *The perception of learning to read as a goal* is evidenced by such behavior as the pupil asking the teacher for reading activities, participating voluntarily in reading activities, choosing reading activities rather than others.

Postulates

1. When reading is learned by means of the sequence: written word presentation, vocal response by the teacher, vocal response by the pupil, the frequency of occurrence of this sequence is related to the accuracy of response of the pupil. (Reader, note that this method of learning to read is commonly referred to as the "look-and-say method" and will be so referred to here.)
2. The effectiveness of the look-and-say method in generating correct reading responses in the pupil is related to the ability of the pupil to discriminate form and shape. Pupils must have a minimum of the latter ability if the method is to produce learning. Additional increments of the ability beyond the minimum result in increased rates of learning.
3. The effectiveness of the look-and-say method in producing correct reading responses is related to the extent to which the pupil perceives the learning of reading as a desirable goal and is motivated to achieve that goal.

Deductions

1. Measures of motivation to read will be correlated with accuracy of response in the early stages of reading in the case of those pupils who perceive reading as a desirable goal.
2. Failure to discriminate two words is a function of the similarity of the shape of the two words.
3. The look-and-say method produces greater accuracy of response when it is supplemented by procedures that emphasize the discrimination of the form of one word and the form of another than it does when such methods are not used.

Note that the deductions do not follow directly from the more comprehensive propositions. For example, the second deduction, while generally related to the second proposition, is also related to hypotheses about the relationship between similarity of forms and difficulty in discrimination. This deductive hypothesis, as stated, is consistent with the second proposition only if we also assume as relatively valid a generalization about the relationship between similarity and discrimination. This hypothesis is deductive because it can be deduced (with additional premises) from the original proposition.

Deductive Hypotheses and the Systematic Development of Knowledge. A little thought will suggest that deductive hypotheses can be particularly useful if we have adequate generalizations from which to make them. One of the purposes of studying educational psychology is to acquire generalizations about human behavior which have relative validity. While a study of educational psychology does

not supply all the generalizations needed for developing a complete set of deductive hypotheses about educational practice, it can supply knowledge sufficient for the development of a good many specific hypotheses.

Corey reports an investigation conducted by a group of social studies teachers, which we can use to illustrate the point that generalizations are available that can serve as a starting point for developing specific hypotheses.³ This group of teachers was interested in determining to what extent a study of biographies of famous figures in American history influences student behavior. The teachers formulated the following hypotheses: (1) that a relationship would be found between the amount of pupil information about these famous Americans and the extent to which they were admired by the pupils; (2) that the degree of admiration for these famous Americans would be increased by one semester of instruction in American history; (3) that a relationship would appear between the traits pupils admired in these historical persons and the reputation that pupils gained for behaving in a manner consistent with the traits they admired.

Our discussion of the acquisition of attitudes should have suggested that the first hypothesis is not likely to be valid. When we analyzed the processes that influence attitude acquisition, we hypothesized that the identification process was crucial in attitude formation. In noting that identification relations probably were not likely if an individual did not know something about the person with whom he might identify, we observed that the process of identification did not depend entirely upon accurate and reliable information about the identification figure.

Should these teachers have formulated this hypothesis? Probably not, unless they were interested in demonstrating again that amount of correctness of information, as such, is not highly correlated with degree of admiration.

The second hypothesis can be criticized on similar grounds. We discussed the influence of courses of instruction on attitude change. We cited data suggesting that participation in a course does not

³ See S. M. Corey, *Action Research to Improve School Practices*. New York: Teachers College, Columbia University Press, 1953, pp. 61-70.

necessarily guarantee changes in attitudes. Since this hypothesis ignored the important variables in attitude formation, the teachers again embarked upon a task that was probably a waste of time.

The origins of the third hypothesis are obscure, but the reasoning seems to run something like this: (1) Famous Americans have admirable qualities; (2) if students admire a famous person, they are likely to have traits similar to that person; (3) since these students have admirable traits, they will have a reputation among their fellow students for possessing these traits. Note that each of these premises is in itself a hypothesis in need of testing. The second premise is the only one of the three that bears some relationship to known psychological generalizations. This principle would be consistent with generalizations about the identification process.

These hypotheses are consistent with a common-sense kind of psychology. In the early stages of development of any science, it is customary to test such common-sense propositions, and in this process of analyzing the obvious, more comprehensive kinds of generalizations are usually developed. But considering the present state of psychological knowledge the three hypotheses above are somewhat naïve.

Theory and Practice in the Development of Systematic Knowledge. There appears to be a tendency in educational research repeatedly to evaluate common-sense propositions of this variety. One of the functions of a course in educational psychology is to provide the teacher with information sufficient for making critical analyses of such obvious generalizations. We hope that after completing this course a student will be able to formulate more sophisticated hypotheses about personality and behavior change than the ones mentioned above.

On the other hand, we do not wish to suggest that the present body of psychological knowledge provides generalizations applicable to each and every educational practice with which teachers are concerned. Many psychological generalizations have been developed in laboratory situations which do not entirely duplicate classroom conditions; they may even have been developed from experimentation with lower animals. One of the functions of educational research is to test the validity of these generalizations in the educational context. Even if psychological science were more completely developed

than it is at the present time, we would still need to evaluate generalizations and their derivations continually.

In many cases, teachers can participate in the development of systematic knowledge by active cooperation in research programs and by suggesting hypotheses which, from their own experience, need testing. In short, we are proposing a more intimate relation between theory and practice. When suggestions for practice are developed from comprehensive generalizations, they need to be tested in classroom situations. Similarly, suggestions for fruitful hypotheses can be arrived at by careful, critical evaluations of classroom experience. In this way theory is nourished by two roots: (1) by new derivations from known generalizations and (2) by the more comprehensive generalizations which can be made by induction from specific practices.

Defining Variables To Be Tested. Since a hypothesis relates two variables, it cannot be tested until its variables are precisely defined in terms of observable behavior. Many variables in hypotheses are formulated at such a high level of abstraction that the hypotheses cannot be directly tested. We have noted this earlier, but it bears repeating. Consider the hypothesis: "Effective leadership enhances group morale." Before this can be tested, we must define "effective leadership." And since the process of definition requires a restatement of the variable in terms of observable behaviors, we must answer the question, "How can we recognize 'effective leadership'?"

The same problem exists for the variable "morale." Think of a class of students for the moment. What would we mean if we said that a class had "morale"? Would we mean that the students will do what the teacher tells them? Or would we mean that the members of the class like each other? Would we mean that the members will work together to attain common goals? Even after choosing any one of these as definitions for the variable, we would still need to define the behaviors we recognize as "working for a common goal," "liking each other," and "doing what the teacher tells them."

From the viewpoint of the formulator of a hypothesis, a variable is a concept for discriminating aspects of observable events. In the first place, the concept of the variable must be clearly defined. Moreover, adequate terminology needs to be developed which conveys the concept in the meaning that the hypothesis-maker intends.

One of the problems in developing hypotheses in professional education is that the concepts and variables we discuss have not been adequately defined. Indeed, we do not have an appropriate technical language for describing concepts. We tend, rather, to use everyday language, with the result that our concepts frequently connote more than we intend. Consider, for example, the variable called "permissiveness." Every teacher has heard a wide variety of generalizations about the relationship of permissiveness to changes in student behavior. But what is "permissiveness"? The concept of the variable itself has not been adequately defined in terms of observable behavior; accordingly, we seldom find a detailed and specific description of a "permissive teacher," or a "permissive classroom climate." Furthermore, the word "permissiveness" itself probably has more connotations than were originally intended for it; this being so, teachers can fill the vacant, undefined concept of the variable with their own individual understandings of the term "permissiveness."

Some terms may actually encourage false interpretation. Coladarci and Getzels have noted this danger in the use of metaphorical language:

A word should be said about the metaphoric flavor of pedagogical language. Metaphors and similes are useful devices in communication. They may become dangerous, however, when the analog unconsciously is translated into a reality. It may have been useful, for instance, for a resident of early Salem to note that "Dame Robinson acts like a witch"—if we assume that what is meant is that the behavior of Dame Robinson resembles that represented in the stereotype of fictional witches. It becomes dangerous only when one is led to think of Dame Robinson *as* a witch and to treat her accordingly—when actually the treatment for Dame Robinson should be quite different from the treatment of witches. That is, one may be led to talking about witches rather than about Dame Robinson. . . . We must be sensitive to the possibility that such language will lead to asking the wrong questions and defining the wrong problems.⁴

Even some terms which have precise denotations in the social sciences carry with them the more vague connotations they have in

⁴ A. P. Coladarci and J. W. Getzels, *The Use of Theory in Educational Administration*. Educational Administration Monograph No. 5, Stanford, California, Stanford University Press, 1955, p. 13.

common, everyday usage. For example, a teacher may describe a child as one who is "maladjusted." What does the teacher mean by the term "maladjusted"? The teacher may be using the word in the general sense, and the content of his communication about the child will be less clear since he has not precisely defined what he means by "maladjusted." We have discussed this problem in other contexts throughout this book; here we are mainly concerned with pointing out that only clear definitions of variables can enable us to formulate testable hypotheses.

In the last several sections, we have outlined principles for formulating testable hypotheses. These principles can be used by any classroom teacher; any teacher should be able to ask clear questions and make meaningful statements about educational practice—indeed, the purpose of this book is to help him do so.

Testing Hypotheses. Hypotheses must be tested as well as formulated. As we have suggested, the formal testing of a hypothesis requires a knowledge of experimental design and technical skills that classroom teachers seldom have. We hope that some teachers will be stimulated to acquire these technical skills. But even if a substantial number of teachers did learn the procedures of experimental research, most of them would not have time to conduct formal research. Does this mean that teachers are unable to evaluate their hypotheses about educational practice?

Some hypotheses may be simple enough for the teacher to conduct comparatively systematic evaluations of them. Others may be more complex, requiring the use of more formal procedures. Even in the latter case, however, teachers can begin the evaluation process by observing effects. Suppose, for example, that a teacher predicts that a new spelling method will work more effectively. A strict test of this hypothesis would require that we set up two comparable groups of students, one of which learns by the new spelling method and the other by the old method; we would first test the spelling performance of both groups under the old method, and then, after a period of time during which one group is taught by the old method and the other by the new one, we would compare the performance of the two groups. This design would provide a reasonably careful test of the hypothesis.

Although the teacher may not be able to implement such a design, he can observe the effects of using a new method. While using a new method, he can ask: Have the pupils improved? Are they learning to spell quickly and easily? Do they seem to be interested in spelling?

We are not suggesting simply that teachers try something "to see if it works." We are suggesting that teachers attempt to gather data systematically *to find out if predicted effects do occur*. Gathering and using factual data on behavior change eliminates some of the inaccuracies inherent in general impressions of student performance. A teacher can keep a record of the "success" and "failures" of an educational practice. With such a record, he is better equipped to ferret out possible reasons for the relative success or failure of his procedures, and to revise and retest them accordingly.

DAVID P. AUSUBEL

Bureau of Educational Research, University of Illinois

Viewpoints from Related Disciplines: Human Growth and Development*

In the past the relationship between psychology and education has often been one of wholesale borrowing by educators from clinical and developmental psychology. The field of human growth and development is, Ausubel believes, a *pure science* and, as such, is in no position to endorse particular educational practices or supply principles for curriculum development. The need for research at the "engineering" level is as important here as it is in learning psychology and education in general (see Melton, pp. 24-25). Furthermore, the concept of *maturation*, that is, those growth processes controlled by

* Reprinted and abridged with the permission of the author and the publisher from the article of the same title, *Teachers College Record*, 60 (1959), 245-254.

hereditary factors, have often been erroneously equated with the concept of *readiness*, that is, those conditions of learning which can be controlled environmentally (see Gagné and Bolles, pp. 36-41). When teachers have confused maturation with a process of "internal ripening," happening irrespective of environment, parents, and teachers, the only practical course left to the teacher is to remain out of the child's way and permit him to "grow" without the corrupting influences of school and civilization, like Rousseau's noble savage.

Ausubel indicates some areas of educational research which could employ developmental concepts, areas that would give the concepts the specificity which McDonald has described as necessary before teachers can use them in classroom practice. He refers particularly to the area of cognitive development. The student may compare what Ausubel suggests here with the approach of Gagné and Bolles to see how developmental and learning psychologists differ in their approach to the same problem. Ausubel's view of motivation as primarily externally derived will be fully discussed by Harlow and illustrated by reports on experimental studies on motivation in the next chapter.

In general, the student should find this article refreshing in the help it furnishes in distinguishing misapplications of psychological concepts and principles and in opening the door to an experimental and genuinely reality-testing approach to the improvement of educational practice.

Child Development and Educational Practice

What light can the field of human growth and development throw on the issue "What shall the schools teach?" I only wish it were possible for me to list and discuss a dozen or more instances in which developmental principles have been validly utilized in providing definitive answers to questions dealing with the content and organization of the curriculum. Unfortunately, however, it must be admitted that at present our discipline can offer only a limited number of very crude generalizations and highly tentative suggestions

bearing on this issue. In a very general sense, of course, it is undeniable that concern with child development has had a salutary effect on the educational enterprise. It alerted school administrators to the fact that certain minimal levels of intellectual maturity were necessary before various subjects could be taught with a reasonable degree of efficiency and hope of success; and it encouraged teachers in presenting their subject matter to make use of the existing interests of pupils, to consider their point of view, and to take into account prevailing limitations in command of language and grasp of concepts. On the other hand, premature and wholesale extension of developmental principles to educational theory and practice has caused incalculable harm. It will take at least a generation for teachers to unlearn some of the more fallacious and dangerous of these overgeneralized and unwarranted applications.

Much of the aforementioned difficulty proceeds from failure to appreciate that human growth and development is a pure rather than an applied science. As a pure science it is concerned with the discovery of general laws about the nature and regulation of human development *as an end in itself*. Ultimately, of course, these laws have self-evident implications for the realization of practical goals in such fields as education, child rearing, and guidance. In a very general sense they indicate the effects of different interpersonal and social climates on personality development and the kinds of methods and subject-matter content that are most compatible with developmental capacity and mode of functioning at a given stage of growth. Thus, because it offers important insights about the changing intellectual and emotional capacities of children as developing human beings, child development may legitimately be considered one of the basic sciences underlying education and guidance and as part of the necessary professional preparation of teachers—in much the same sense that anatomy and bacteriology are basic sciences for medicine and surgery.

Actual application to practical problems of teaching and curriculum, however, is quite another matter. Before the educational implications of developmental findings can become explicitly useful in everyday school situations, much *additional* research at the engineering level of operations is necessary. Knowledge about nuclear fission, for example, does not tell us how to make an atomic bomb

or an atomic-powered submarine, antibiotic reactions that take place in petri dishes do not necessarily take place in living systems, and methods of learning employed by animals in mazes do not necessarily correspond to methods of learning that children use in grappling with verbal materials in classrooms. Many of the better-known generalizations in child development—the principle of readiness, the cephalocaudal trend, the abstract to concrete trend in conceptualizing the environment, and others—fit these analogies perfectly. They are interesting and potentially useful ideas to curriculum specialists but will have little practical utility in designing a social studies or physical education curriculum unless they are rendered more specific in terms of the actual operations involved in teaching these subjects. This lack of fruitful particularization, although unfortunate and regrettable, does not in itself give rise to damaging consequences except insofar as many beginning teachers tend to nurture vague illusions about the current usefulness of these principles, and subsequently, after undergoing acute disillusionment, lose the confidence they may have felt in the value of a developmental approach to educational problems.

Keeping these qualifications about the relevance of child development for educational practice in mind, I propose briefly to consider from the standpoint of developmental psychology the following aspects of the issue under discussion: (1) readiness as a criterion for curricular placement; (2) developmental factors affecting breadth of the curriculum; (3) the child's voice in determining the curriculum; and (4) the content and goals of instruction in relation to the organization and growth of the intellect.

Readiness and Grade Placement

There is little disagreement about the fact that readiness always crucially influences the efficiency of the learning process and often determines whether a given intellectual skill or type of school material is learnable at all at a particular stage of development. Most educators implicitly accept also the proposition that an *optimal* age exists for every kind of learning. Postponement of learning experience beyond the age of optimal readiness wastes valuable and often unsuspected learning opportunities, thereby unnecessarily reducing

the amount and complexity of subject matter content that can be mastered in a designated period of schooling. It is also conceivable that beyond a certain critical age the learning of various intellectual skills becomes more difficult for an older than for a younger child. On the other hand, when a pupil is prematurely exposed to a learning task before he is ready for it, he not only fails to learn the task in question but even learns from the experience of failure to fear, dislike, and avoid it.

Up to this point, the principle of readiness—the idea that attained capacity limits and influences an individual's ability to profit from current experience or practice—is empirically demonstrable and conceptually unambiguous. Difficulty first arises when it is confused with the concept of *maturation* and when the latter concept in turn is equated with a process of “internal ripening.” The concept of readiness simply refers to the adequacy of existing capacity in relation to the demands of a given learning task. No specification is made as to *how* this capacity is achieved—whether through prior practice of a specific nature (learning), through incidental experience, through genically regulated structural and functional changes occurring independently of environmental influences, or through various combinations of these factors. Maturation, on the other hand, has a different and much more restricted meaning. It encompasses those increments in capacity that take place in the demonstrable absence of specific practice experience—those that are attributable to genic influences and/or incidental experience. Maturation, therefore, is not the same as readiness but is merely one of the two principal factors (the other being learning) that contribute to or determine the organism's readiness to cope with new experience. Whether or not readiness exists, in other words, does not necessarily depend on maturation alone but in many instances is solely a function of prior learning experience and most typically depends on varying proportions of maturation and learning.

To equate the principles of readiness and maturation not only muddies the conceptual waters but also makes it difficult for the school to appreciate that insufficient readiness may reflect inadequate prior learning on the part of pupils because of inappropriate or inefficient instructional methods. Lack of maturation can thus become a convenient scapegoat whenever children manifest insuf-

ficient readiness to learn, and the school, which is thereby automatically absolved of all responsibility in the matter, consequently fails to subject its instructional practices to the degree of self-critical scrutiny necessary for continued educational progress. In short, while it is important to appreciate that the current readiness of pupils determines the school's current choice of instructional methods and materials, it is equally important to bear in mind that this readiness itself is partly determined by the appropriateness and efficiency of the previous instructional practices to which they have been subjected.

The conceptual confusion is further compounded when maturation is interpreted as a process of "internal ripening" essentially independent of *all* environmental influences, that is, of both specific practice and incidental experience. Readiness then becomes a matter of simple genic regulation unfolding in accordance with a predetermined and immutable timetable; and the school, by definition, becomes powerless to influence readiness either through its particular way of arranging specific learning experiences or through a more general program of providing incidental or nonspecific background experience preparatory to the introduction of more formal academic activities.

Actually, the embryological model of development implicit in the "internal ripening" thesis fits quite well when applied to human sensorimotor and neuromuscular sequences taking place during the prenatal period and early infancy. In the acquisition of simple behavioral functions (for example, locomotion, prehension) that characterize all members of the human species irrespective of cultural or other environmental differences, it is reasonable to suppose that for all practical purposes genic factors alone determine the direction of development. Environmental factors only enter the picture if they are extremely deviant, and then serve more to disrupt or arrest the ongoing course of development than to generate distinctive developmental progressions of their own. Thus, the only truly objectionable aspect of this point of view is its unwarranted extrapolation to those more complex and variable components of later cognitive and behavioral development where unique factors of individual experience and cultural environment make important contributions to the direction, patterning and sequential order of all developmental changes.

It is hardly surprising, therefore, in view of the tremendous influence on professional and lay opinion wielded by Gesell and his colleagues, that many people conceive of readiness in absolute and immutable terms, and thus fail to appreciate that except for such traits as walking and grasping, the mean ages of readiness can never be specified apart from relevant environmental conditions. Although the modal child in contemporary America may first be ready to read at the age of six and one-half, the age of reading readiness is always influenced by cultural, subcultural, and individual differences in background experience, and in any case varies with the method of instruction employed and the child's IQ. Middle-class children, for example, are ready to read at an earlier age than lower-class children because of the greater availability of books in the home and because they are read to and taken places more frequently.

The need for particularizing developmental generalizations before they can become useful in educational practice is nowhere more glaringly evident than in the field of readiness. At present we can only speculate what curricular sequences might conceivably be if they took into account precise and detailed (but currently unavailable) research findings on the emergence of readiness for different subject-matter areas, for different sub-areas and levels of difficulty within an area, and for different techniques of teaching the same material. Because of the unpredictable specificity of readiness as shown, for example, by the fact that four- and five-year-olds can profit from training in pitch but not in rhythm, valid answers to such questions cannot be derived from logical extrapolation but require meticulous empirical research in a school setting. The next step would involve the development of appropriate teaching methods and materials to take optimal advantage of existing degrees of readiness and to increase readiness wherever necessary and desirable. But since we generally do not have this type of research data available, except perhaps in the field of reading, we can only pay lip service to the principle of readiness in curriculum planning.

Breadth of Curriculum

One of the chief complaints of the critics of public education, both in the United States and in New Zealand, is that modern children fail to learn the fundamentals because of the broadening of the ele-

mentary school curriculum to include such subjects as social studies, art, science, music, and manual arts in addition to the traditional three R's. This, of course, would be a very serious charge if it were true, because the wisdom of expanding a child's intellectual horizons at the expense of making him a cripple in the basic intellectual skills is highly questionable to say the least. Fortunately, however, the benefits of an expanded curriculum have thus far not been accompanied by a corresponding deterioration in the standard of the three R's. Evidently the decreased amount of time spent on the latter subjects has been more than compensated for by the development of more efficient methods of teaching and by the incidental learning of the fundamentals in the course of studying these other subjects. Nevertheless, the issue of breadth versus depth still remains because there is obviously a point beyond which increased breadth could only be attained by sacrificing mastery of the fundamental skills; and even if we agreed to maintain or improve the present standard of the three R's, we would still have to choose between breadth and depth in relation to other components of the curriculum, particularly at the junior and senior high school levels. It is at these points of choice that developmental criteria can be profitably applied.

Generally speaking, maximal breadth of the curriculum consistent with adequate mastery of its constituent parts is developmentally desirable at all ages because of the tremendously wide scope of human abilities. The wider the range of intellectual stimulation to which pupils are exposed, the greater are the chances that all of the diverse potentialities both within a group of children and within a single child will be brought to fruition. By the same token, a broad curriculum makes it possible for more pupils to experience success in the performance of school activities and thus to develop the necessary self-confidence and motivation for continued academic striving and achievement. The very fact that elementary school children are able to make significant progress in science and social studies also indicates that myopic concentration on the three R's would waste much available readiness for these types of learnings and thus compel junior and senior high schools to devote much of their instructional time to materials that are easily learnable in the lower grades. In fact, one of the major failings of the secondary school curriculum today is that because it still has not adequately

adjusted to the expansion of the elementary school syllabus, entering pupils are subjected to much stultifying repetition and fail to break the new ground for which they are obviously ready.

The relationship between breadth and depth must also take into account the progressive differentiation of intelligence, interests, and personality structure with increasing age. The elementary school child is a "generalist" because both his intellect and his personality are still relatively unstable and uncrystallized and lack impressive internal consistency. Thus, many different varieties of subject matter are equally compatible with his interest and ability patterns. Furthermore, unless he has experience with many different fields of knowledge and gives each a provisional try, he is in no position to judge which kinds of intellectual pursuits are most congruent with his major ability and value systems. Hence, quite apart from the future life adjustment values of a broad educational background, it is appropriate on developmental grounds for elementary and early high school curricula to stress breadth rather than depth.

Toward the latter portion of the high school period, however, precisely the opposite kind of situation begins to emerge. Interests have crystallized and abilities have undergone differentiation to the point where greater depth and specialization are possible and desirable. Many students at this stage of intellectual development are ready to sink their teeth into more serious and solid academic fare, but unfortunately suitable instructional programs geared at an advanced level of critical and independent thinking are rarely available. The changes that have taken place in secondary school curricula since the academy days have been primarily characterized by the belated and half-hearted addition of more up-to-date and topical information. Very little has been done in the way of providing the student with a meaningful, integrated, systematic view of the major ideas in a given field of knowledge.

The Child's Voice in Curriculum Planning

One extreme point of view associated with the child-centered approach to education is the notion that children are innately equipped in some mysterious fashion for knowing precisely what is best for them. This idea is obviously an outgrowth of predeterministic theories (for example, those of Rousseau and Gesell) that con-

ceive of development as a series of internally regulated sequential steps that unfold in accordance with a prearranged design. According to these theorists, the environment facilitates development best by providing a maximally permissive field that does not interfere with the predetermined processes of spontaneous maturation. From these assumptions it is but a short step to the claim that the child himself must be in the most strategic position to *know* and *select* those components of the environment that correspond most closely with his current developmental needs and hence are most conducive to optimal growth. Empirical "proof" of this proposition is adduced from the fact that nutrition is adequately maintained and existing deficiency conditions are spontaneously corrected when infants are permitted to select their own diets. If the child can successfully choose his diet, he must certainly know what is best for him in all areas of growth and should therefore be permitted to select everything, including his curriculum.

In the first place, and refuting this theory, even if development were primarily a matter of internal ripening, there would still be no good reason for supposing that the child is therefore implicitly conversant with the current direction and facilitating conditions of development and hence axiomatically equipped to make the most appropriate choices. Because the individual is sensitive in early childhood to internal cues of physiological need we cannot conclude that he is similarly sensitive to cues reflective of psychological and other developmental needs; even in the area of nutrition, selection is a reliable criterion of need only during early infancy.

Second, unless one assigns a sacrosanct status to endogenous motivations, there is little warrant for believing either that they alone are truly reflective of the child's *genuine* developmental requirements or that environmentally derived needs are "imposed," authoritarian in spirit, and inevitably fated to thwart the actualization of his developmental potentialities. Actually, most needs originate from without and are internalized in the course of the child's interaction and identification with significant persons in his family and cultural environments.

Third, one can never assume that the child's *spontaneously* expressed interests and activities are completely reflective of *all* of his important needs and capacities. Just because capacities can poten-

tially provide their own motivation does not mean that they always or necessarily do so. It is not the possession of capacities that is motivating, but the anticipation of future satisfactions once they have been successfully exercised. But because of such factors as inertia, lack of opportunity, lack of appreciation, and preoccupation with other activities, many capacities may never be exercised in the first place. Thus, children typically develop only *some* of their potential capacities, and their expressed interests cannot be considered coextensive with the potential range of interests they are capable of developing with appropriate stimulation.

In conclusion, therefore, the current interests and spontaneous desires of immature pupils can hardly be considered reliable guideposts and adequate substitutes for specialized knowledge and seasoned judgment in designing a curriculum. Recognition of the role of pupil needs in school learning does not mean that the scope of the syllabus should be restricted to the existing concerns and spontaneously expressed interests that happen to be present in a group of children growing up under particular conditions of intellectual and social class stimulation. In fact, one of the primary functions of education should be to stimulate the development of motivations that are currently nonexistent. It is true that academic achievement is greatest when pupils manifest felt needs to acquire knowledge as an end in itself. Such needs, however, are not endogenous but acquired—and largely through exposure to provocative, meaningful, developmentally appropriate instruction. Hence, while it is reasonable to consider the views of pupils and even, under certain circumstances, to solicit their participation in the planning of the curriculum, it makes little developmental or administrative sense to entrust them with responsibility for significant policy or operational decisions. R =

Organization and Cognitive Development

The curriculum specialist is concerned with more than the appropriate grade placement of different subjects and subject-matter content in accordance with such criteria as readiness and relative significance for intellectual, vocational, or current adjustment purposes. More important than what pupils know at the end of the

sixth, eighth, and twelfth grades is the extent of their knowledge at the ages of twenty-five, forty and sixty as well as their ability and desire both to learn more and to apply their knowledge fruitfully in adult life. In light of these latter criteria, in comparing, for example, the quantity and quality of our national research output in the pure and applied sciences with those of European countries, the American educational system stands up relatively well even though our school children apparently absorb less academic material. We are dealing here with the ultimate intellectual objectives of schooling, namely, with the long-term acquisition of stable and usable bodies of knowledge and intellectual skills and with the development of ability to think creatively, systematically, independently, and with depth in particular fields of inquiry. Instruction obviously influences the outcome of these objectives—not so much in the substantive content of subject matter but in the organization, sequence, and manner of presenting learning experiences, their degree of meaningfulness, and the relative balance between conceptual and factual materials.

But obviously, before we could ever hope to structure effectively such instructional variables for the optimal realization of these designated objectives, we would have to know a great deal more about the organizational and developmental principles whereby human beings acquire and retain stable bodies of knowledge and develop the power of critical and productive thinking. This type of knowledge, however, will forever elude us unless we abandon the untenable assumption that there is no real distinction either between the logic of a proposition and how the mind apprehends it or between the logical structure of subject-matter organization and the actual series of cognitive processes through which an immature and developing individual incorporates facts and concepts into a stable body of knowledge. It is perfectly logical from the standpoint of a mature scholar, for example, to write a textbook in which topically homogenous materials are segregated into discrete chapters and treated throughout at a uniform level of conceptualization. But how closely does this approach correspond with highly suggestive findings that one of the major cognitive processes involved in the learning of any new subject is progressive differentiation of an originally undifferentiated field? Once we learn more about cognitive development than the crude generalizations that developmental

psychology can currently offer, it will be possible to employ organizational and sequential principles in the presentation of subject matter that actually parallel developmental changes in the growth and organization of the intellect. In the meantime let us examine briefly how such generalizations as the concrete-to-abstract trend, the importance of meaningfulness, and the principle of retroactive inhibition have been used and abused in educational practice.

Many features of the activity program are based on the premise that the elementary school child perceives the world in relatively specific and concrete terms and requires considerable firsthand experience with diverse concrete instances of a given set of relationships before he can abstract genuinely meaningful concepts. Thus, an attempt is made to teach factual information and intellectual skills in the real-life functional contexts in which they are customarily encountered rather than through the medium of verbal exposition supplemented by artificially contrived drills and exercises. This approach has real merit, if a fetish is not made of naturalism and incidental learning, if drills and exercises are provided in instances where opportunities for acquiring skills do not occur frequently and repetitively enough in more natural settings, and if deliberate or guided effort is not regarded as incompatible with incidental learning. Even more important, however, is the realization that in older children, once a sufficient number of basic concepts are consolidated, new concepts are primarily abstracted from verbal rather than from concrete experience. Hence in secondary school it may be desirable to reverse both the sequence and the relative balance between abstract concepts and supportive data. There is good reason for believing, therefore, that much of the time presently spent in cook-book laboratory exercises in the sciences could be much more advantageously employed in formulating precise definitions, making explicit verbal distinctions between concepts, generalizing from hypothetical situations, and in other ways. ✓

Another underlying assumption of activity and project methods is that concepts and factual data are retained much longer when they are meaningful, genuinely understood, and taught as larger units of interrelated materials than when they are presented as fragmented bits of isolated information and committed to rote memory. This, of course, does not preclude the advisability of rote learning for certain kinds of learning (for example, multiplication tables) after a

functional understanding of the underlying concepts has been acquired. Unfortunately, however, these principles have made relatively few inroads on the high school instructional program, where they are still applicable. The teaching of mathematics and science, for example, still relies heavily on rote learning of formulas and procedural steps, on recognition of traditional "type problems," and on mechanical manipulation of symbols. In the absence of clear and stable concepts which serve as anchoring points and organizing foci for the assimilation of new material, secondary school students are trapped in a morass of confusion and seldom retain rotely memorized materials much beyond final exam time.

This brings us finally to a consideration of the mechanisms of accretion and long-term retention of large bodies of ideational material. Why do high school and university students tend to forget so readily previous day-to-day learnings as they are exposed to new lessons? The traditional answer of educational psychology, based upon studies of short-term rote learning in animal and human subjects, has been that subsequent learning experiences which are similar to but not identical with previously learned materials exert a retroactively inhibitory effect on the retention of the latter. But wouldn't it be reasonable to suppose that all of the existing, cumulatively established ideational systems which an individual brings with him to any learning situation have more of an interfering effect on the retention of new learning material (proactive inhibition) than brief exposure to subsequently introduced materials of a similar nature (retroactive inhibition)? Because it is cognitively most economical and least burdensome for an individual to subsume as much new experience as possible under existing concepts that are inclusive and stable, the import of many specific illustrative items in later experience is assimilated by the generalized meaning of these more firmly established and highly conceptualized subsuming foci. When this happens the latter items lose their identity and are said to be "forgotten." Hence, if proactive rather than retroactive inhibition turned out to be the principal mechanism affecting the longevity with which school materials were retained, it would behoove us to identify those factors that counteract it and to employ such measures in our instructional procedures.

*Motivation:
Does Curiosity
Kill the Cat?*

Introduction

The question "Why do we behave as we do?" has always profoundly concerned mankind. It is a question that has special interest to the experimental psychologist, who, through his knowledge of motivation, hopes to improve his predictions of behavior. And because the teacher and the educational psychologist also hope to improve their prediction and control of student behavior in the classroom, they share the same interest in human motivation.

The question of why we behave as we do has been variously answered. The Greeks attributed the control of human affairs to Fate—a polite defense of the behavior of their irascible and unpredictable gods. Judaism and Christianity saw in a single, alternately condoning and punishing God the controlling factor in human affairs. Both Darwin, with his interest in instinct and adaptation (or learning), and Freud, with his concern with the libido (or pleasure) and the death instinct, made the source of motivation more internal than external, yet, for all of that, not very manageable by man. In fact, most contemporary theories of motivation assert that the internal states or needs of the organism are the basis for behavior; this point of view discussed more fully below, does not, however, receive the assent of all psychologists.

Much of the current experimental literature on motivation continues to emphasize "drive-reduction" (the relieving of certain inner states of physiological tension) as the explanation of human be-

havior. (For example, the contractions of a hungry stomach are relieved after food has been taken.) Motives are regarded as bridges between needs on the one hand and behavior on the other. These needs are often various forms of hungers, for example, thirst or pain. When the needs are seen as the basis for the organism's activity, they are called *drives*. According to the "drive-reduction" theorists, the drive persists until the need is satisfied.*

In drive-reduction theory the satisfaction of physiological (primary) needs leads man into a complicated nexus of social relations. This situation in turn gives rise to a host of secondary needs, motives, and drives. There are many varying classifications of secondary motives, but some examples are social approval, achievement, affiliation, and security. In theory these are learned motivations and, if learned, they could have educational significance. However, as it presently stands, this view offers very little guidance for the organization of learning situations in the classroom, either in the utilization of already existing secondary motives to help achieve educational objectives, or in facilitation by the teacher and the school of the learning of new motives. The development and modification of motivational structures of this nature is a study which may be more profitably pursued in the clinic than in the classroom.

Newer approaches in the study of motivation have taken issue with the view that behavior must always start with some internal deficiency of the organism, especially when this organism is a human being. While it is quite sophisticated to assert that much of our social and literary activity in mid-century America is motivated by sex-hunger, casual observation reveals that sexual activity is often less a need than a means to some other goal. Experimental psychologists are discovering that they can make better predictions of behavior in terms of *goals* and *incentives* than in terms of needs as such. Rewards, and even the stimuli associated with rewards, become a source of motivation. Such aspects of behavior as ego-involvement and perceived value can serve as motivational sources. With a surplus of goals, incentives, rewards, etc., to choose from, it would

* See John F. Hall, *Psychology of Motivation* (Philadelphia: J. B. Lippincott Company, 1961), pp. 30-36.

seem that much human motivation can be as well explained by an abundance of stimulation as by inner deficiencies.

Thus recent motivational theory emphasizes concepts of "pull" rather than "push." A major departure from drive-reduction theory has been the work of experimental psychologists who are attempting to discover how much external stimulation arouses organisms to activity, especially human beings. The idea that external stimulation is a major source of motivation appeals to the teacher and the educational psychologist because it offers more hope of managing the motivational aspect of all learning situations than does the older theory. The work of Harlow and Terrell on the "manipulation drive" (see pp. 85-96 and pp. 98-106), of K. C. Montgomery and others on the "exploratory drive," and of Canadian and American psychologists on "sensory deprivation" all reflect current experimental attempts to explain motivation by external stimulation rather than by internal states of need and drive.

Although educational psychologists recognize motivation as a key aspect of all learning situations, it has been difficult to translate the results of laboratory experimentation into working hypotheses for classroom research. In the first place, it is obvious that, especially for middle-class children, the need for food, for example, is hardly going to get well-fed Mary and over-stuffed Johnny to improve their reading. As we move up the phylogenetic scale (see Harlow, pp. 93-94), it would seem that, especially for monkeys and people, the attractiveness of the environment (the simple "need" for the eyes to see and the ears to hear) and the social valuation placed on environmental stimuli (for example, money) all become plausible explanations for what triggers activity, even though the experimental support for such explanations is hard to find. In fact, motivation may be an area in which the educational psychologists, by focusing on human rather than animal experimentation, may contribute more to the experimental psychologists than they borrow from them.

Bugelski, who has stated that experimental psychologists in their work on motivation have not freed themselves from the linguistic fiction of a previous historical period, suggests that the question of motivation in teaching may boil down to getting the student's attention. He states:

The primary function of the teacher, then, is to get the student to "pay" attention. Whether this is done by use of birch rods, electric shock, Chinese water torture, or by promises of ice cream, movies, or money is irrelevant. One procedure or another may involve disastrous side effects or lead to difficulties, but if attention is secured that problem is solved. Sometimes even screaming at the children to "Pay Attention!" may work.*

Even this modest conception of the role of motivation in teaching directs attention to external stimulation (of sorts) rather than to the inner states of the student. The demonstration of the effect of *set* on motivation and learning in the study by Wittrock may partially support Bugelski's point (pp. 107-117).

There are other aspects of motivation which have attracted the attention of experimental psychologists. The effect of anxiety and frustration on the quality of performance is another case of interest. The study of Child and Waterhouse can illustrate what psychologists are doing here (see pp. 117-128).

Relationship of Readings in Chapter Two

The selection by Harlow is an interesting defense of the point of view on motivation which is emphasized in this book of readings, that the chief source of motivation is external stimulation. This statement is made simply as a proposition which awaits further testing in the laboratory and the classroom and is in no way a statement of dogma. The Harlow article establishes the frame of reference for the study by Terrell on the "manipulatory drive" in children.

The three remaining readings deal with other aspects of motivation, but they furnish further evidence that needs and drives are not the only sources of motivation. The study by Wittrock attempts to discover how *set* (as a source of motivation) influences student achievement. The study by Child and Waterhouse concerns the motivating effect of *frustration* when combined with ego-involvement. The last study, by Bostrom, examines the effect of grades on subsequent student performance and provides an example of how rewards act as a motivational source.

* R. B. Bugelski, *Psychology of Learning* (New York: Holt, Rinehart and Winston, Inc., 1961), p. 481.

With the exception of the selection by Harlow, all the readings in this chapter are direct research reports of experimentation with human subjects. By observing the *methods* as well as the results of this research, the education student can gain some understanding of the experimental method which McDonald described in Chapter 1 (pp. 52-67) as a means for improving classroom practice.

HARRY F. HARLOW
University of Wisconsin

Mice, Monkeys, Men, and Motives*

In this amusing and important article Harlow asserts that one reason why we learn or do anything is our tendency to be inquisitive about and playful with an environment which manages to keep us fairly well aroused much of the time. In their discussion of motivation, Gagné and Bolles also have referred to the human tendency to gamble and to the interest the learners often show in the task itself (see pp. 36-37).

In defense of his position, Harlow points to the limitations of the drive-reduction theory of motivation (see pp. 90-94) and to the evidence from his own experimental studies. The drive-reduction theory is based mostly on experimentation with animals such as the rat. Since these animals are fairly low on the evolutionary scale, the results of experiments on them may be of minor importance in explaining the complexities of human behavior. Here Harlow echoes the complaint of Melton, that more experimental work should be done with children (see pp. 6-7). Harlow also reports observations made of the new-born human infant, which responds immediately to touch (an example of environmental stimulation) without learning or

* Reprinted and abridged with the permission of the author and the American Psychological Association from the article of the same title, *Psychological Review*, 60 (1953), 23-32.

"needing" to. He reminds us also that our best learning does not occur under extreme drive conditions: we are not most curious when we are most hungry—except about food only. He refers to his own experimental work with monkeys, who, he contends, have never heard of the drive-reduction theory and therefore seem to ignore it in their behavior.

Although avoiding some of the limitations of the older theory, Harlow's theory of motivation offers difficulties of its own. For example, he speaks of a manipulation drive, but, as has been explained above, drives are based on needs. When we speak of a sex-drive, we know the source is a need for sexual satisfaction. But when a monkey works a puzzle, to what precisely in this situation do we attribute his manipulation drive? Also, it is not clear what the monkey finds sufficiently rewarding in such a pastime to sustain his interest and prompt him to return. Some psychologists have stated that "capacity is its own motivation."^{*}

The education student may well consider what testable hypotheses Harlow's view of motivation can furnish the classroom. In considering students' "needs" should we include his "need" to be intellectually curious and exploratory? Does this view suggest that teachers should follow the spontaneously expressed interests of the students in determining what to teach them, a practice associated with progressive education? Or does it suggest, as Ausubel has (pp. 76-77), that as teachers we should constantly expose students to learning environments which they are not yet aware of and which might stimulate new curiosity?

Many of psychology's theoretical growing pains—or, in modern terminology, conditioned anxieties—stem from the behavior revolution of Watson. The new psychology intuitively disposed of instincts and painlessly disposed of hedonism. But having completed

^{*} John F. Hall, *Psychology of Motivation* (Philadelphia: J. B. Lippincott Company, 1961), pp. 96-98.

this St. Bartholomew-type massacre, behavioristic motivation theory was left with an aching void, a nonhedonistic aching void, needless to say.

Before the advent of the Watsonian scourge the importance of external stimuli as motivating forces was well recognized. Psychologists will always remain indebted to Loeb's brilliant formulation of tropistic theory, which emphasized, and probably overemphasized, the powerful role of external stimulation as the primary motivating agency in animal behavior. Unfortunately, Loeb's premature efforts to reduce all behavior to overly simple mathematical formulation, his continuous acceptance of new tropistic constructs in an effort to account for any aberrant behavior not easily integrated into his original system, and his abortive attempt to encompass all behavior into a miniature theoretical system doubtless led many investigators to underestimate the value of his experimental contributions.

Thorndike was simultaneously giving proper emphasis to the role of external stimulation as a motivating force in learning and learned performances. Regrettably, these motivating processes were defined in terms of pain and pleasure, and it is probably best for us to dispense with such lax, ill-defined, subjective terms as pain, pleasure, anxiety, frustration, and hypotheses—particularly in descriptive and theoretical rodentology.

Instinct theory, for all its terminological limitations, put proper emphasis on the motivating power of external stimuli; for, as so brilliantly described by Watson in 1941, the instinctive response was elicited by "serial stimulation," much of which was serial external stimulation.

The almost countless researches on tropisms and instincts might well have been expanded to form a solid and adequate motivational theory for psychology—a theory with a proper emphasis on the role of external stimulus and an emphasis on the importance of incentives as opposed to internal drives *per se*.

It is somewhat difficult to understand how this vast and valuable literature was to become so completely obscured and how the importance of the external stimulus as a motivating agent was to become lost. Pain-pleasure theory was discarded because the terminology had subjective, philosophical implications. Instinct theory fell into disfavor because psychologists rejected the dichotomized

heredity-environment controversy and, also, because the term "instinct" had more than one meaning. Why tropistic theory disappeared remains a mystery, particularly inasmuch as most of the researches were carried out on subprimate animal forms.

Modern motivation theory apparently evolved from an overpopularization of certain experimental and theoretical materials. Jennings' demonstration that "physiological state" played a role in determining the behavior of the lower animal was given exaggerated importance and emphasis, thereby relegating the role of external stimulation to a secondary position as a force in motivation. The outstanding work in the area of motivation between 1920 and 1930 related to visceral drives and drive cycles and was popularized by Richter's idealized theoretical paper on "Animal Behavior and Internal Drives" and Cannon's *The Wisdom of the Body*.

When the self-conscious behavior theorists of the early thirties looked for a motivation theory to integrate with their developing learning constructs, it was only natural that they should choose the available tissue-tension hypotheses. Enthusiastically and uncritically the S-R theorists swallowed these theses whole. For fifteen years they have tried to digest them, and it is now time that these theses be subjected to critical examination, analysis, and evaluation. We do not question that these theses have fertilized the field of learning, but we do question that the plants that have developed are those that will survive the test of time.

It is my belief that the theory which describes learning as dependent upon drive reduction is false, that internal drive as such is a variable of little importance to learning, and that this small importance steadily decreases as we ascend the phyletic scale and as we investigate learning problems of progressive complexity. Finally, it is my position that drive-reduction theory orients learning psychologists to attack problems of limited importance and to ignore the fields of research that might lead us in some foreseeable future time to evolve a theoretical psychology of learning that transcends any single species or order.

There can be no doubt that the single-celled organisms such as the amoeba and the paramecium are motivated to action both by external and internal stimuli. The motivation by external stimulation gives rise to heliotropisms, chemotropisms, and rheotropisms.

The motivation by internal stimulation produces characteristic physiological states which have, in turn, been described as chemotropisms. From a phylogenetic point of view, moreover, neither type of motive appears to be more basic or more fundamental than the other. Both types are found in the simplest known animals and function in interactive, rather than in dominant-subordinate, roles.

Studies of fetal responses in animals from opossum to man give no evidence suggesting that the motivation of physiological states precedes that of external incentives. Tactual, thermal, and even auditory and visual stimuli elicit complex patterns of behavior in the fetal guinea pig, although this animal has a placental circulation which should guarantee against thirst or hunger. The newborn opossum climbs up the belly of the female and into the pouch, apparently in response to external cues; if visceral motives play any essential role, it is yet to be described. The human fetus responds to external tactual and nociceptive stimuli at a developmental period preceding demonstrated hunger or thirst motivation. Certainly, there is no experimental literature to indicate that internal drives are ontogenetically more basic than exteroceptive motivating agencies.

Tactual stimulation, particularly of the cheeks and lips, elicits mouth, head, and neck responses in the human neonate, and there are no data demonstrating that these responses are conditioned, or even dependent, upon physiological drive states. Hunger appears to lower the threshold for these responses to tactual stimuli. Indeed, the main role of the primary drive seems to be one of altering the threshold for precurrent responses. Differentiated sucking response patterns have been demonstrated to quantitatively varied thermal and chemical stimuli in the infant only hours of age, and there is, again, no reason to believe that the differentiation could have resulted from antecedent tissue-tension reduction states. Taste and temperature sensations induced by the temperature and chemical composition of the liquids seem adequate to account for the responses.

There is neither phylogenetic nor ontogenetic evidence that drive states elicit more fundamental and basic response patterns than do external stimuli; nor is there basis for the belief that precurrent

responses are more dependent upon consummatory responses than are consummatory responses dependent upon precurrent responses. There is no evidence that the differentiation of the innate precurrent responses is more greatly influenced by tissue-tension reduction than are the temporal ordering and intensity of consummatory responses influenced by conditions of external stimulation.

There are logical reasons why a drive-reduction theory of learning, a theory which emphasizes the role of internal, physiological-state motivation, is entirely untenable as a motivational theory of learning. The internal drives are cyclical and operate, certainly at any effective level of intensity, for only a brief fraction of any organism's waking life. The classical hunger drive physiologically defined ceases almost as soon as food—or nonfood—is ingested. This, as far as we know, is the only case in which a single swallow portends anything of importance. The temporal brevity of operation of the internal drive states obviously offers a minimal opportunity for conditioning and a maximal opportunity for extinction. The human being, at least in the continental United States, may go for days or even years without ever experiencing true hunger or thirst. If his complex conditioned responses were dependent upon primary drive reduction, one would expect him to regress rapidly to a state of tuitional oblivion. There are, of course, certain recurrent physiological drive states that are maintained in the adult. But the studies of Kinsey indicate that in the case of one of these there is an inverse correlation between presumed drive strength and scope and breadth of learning, and in spite of the alleged reading habits of the American public, it is hard to believe that the other is our major source of intellectual support. Any assumption that derived drives or motives can account for learning in the absence of primary drive reduction puts an undue emphasis on the strength and permanence of derived drives, at least in subhuman animals. Experimental studies to date indicate that most derived drives and second-order conditioned responses rapidly extinguish when the rewards which theoretically reduce the primary drives are withheld. The additional hypothesis of functional autonomy of motives, which could bridge the gap, is yet to be demonstrated experimentally.

The condition of strong drive is inimical to all but very limited aspects of learning—the learning of ways to reduce the internal ten-

sion. The hungry child screams, closes his eyes, and is apparently oblivious to most of his environment. During this state he eliminates response to those aspects of his environment around which all his important learned behaviors will be based. The hungry child is a most incurious child, but after he has eaten and become thoroughly sated, his curiosity and all the learned responses associated with his curiosity take place. If this learning is conditioned to an internal drive state, we must assume it is the resultant of backward conditioning. If we wish to hypothesize that backward conditioning is dominant over forward conditioning in the infant, it might be possible to reconcile fact with S-R theory. It would appear, however, that alternate theoretical possibilities should be explored before the infantile backward conditioning hypothesis is accepted.

Observations and experiments on monkeys convinced us that there was as much evidence to indicate that a strong drive state inhibits learning as to indicate that it facilitates learning. It was the speaker's feeling that monkeys learned most efficiently if they were given food before testing, and as a result, the speaker routinely fed his subjects before every training session. The rhesus monkey is equipped with enormous cheek pouches, and consequently many subjects would begin the educational process with a rich store of incentives crammed into the buccal cavity. When the monkey made a correct response, it would add a raisin to the buccal storehouse and swallow a little previously munched food. Following an incorrect response, the monkey would also swallow a little stored food. Thus, both correct and incorrect responses invariably resulted in S-R theory drive reduction. It is obvious that under these conditions the monkey cannot learn, but the present speaker developed an understandable skepticism of this hypothesis when the monkeys stubbornly persisted in learning, learning rapidly, and learning problems of great complexity. Because food was continuously available in the monkey's mouth, an explanation in terms of differential fractional anticipatory goal responses did not appear attractive. It would seem that the Lord was simply unaware of drive-reduction learning theory when he created, or permitted the gradual evolution of, the rhesus monkey.

The langurs are monkeys that belong to the only family of primates with sacculated stomachs. There would appear to be no

mechanism better designed than the sacculated stomach to induce automatically prolonged delay of reinforcement defined in terms of homeostatic drive reduction. Langurs should, therefore, learn with great difficulty. But a team of Wisconsin students has discovered that the langurs in the San Diego Zoo learn at a high level of monkey efficiency. There is, of course, the alternative explanation that the inhibition of hunger contractions in multiple stomachs is more reinforcing than the inhibition of hunger contractions in one. Perhaps the quantification of the gastric variable will open up great new vistas of research.

Actually, the anatomical variable of diversity of alimentary mechanisms is essentially uncorrelated with learning to food incentives by monkeys and suggests that learning efficiency is far better related to tensions in the brain than in the belly.

Experimental test bears out the fact that learning performance by the monkey is unrelated to the theoretical intensity of the hunger drive. Meyer tested rhesus monkeys on discrimination-learning problems under conditions of maintenance-food deprivation of 1.5, 18.5, and 22.5 hours and found no significant differences in learning or performance. Subsequently, he tested the same monkeys on discrimination-reversal learning following 1, 23, and 47 hours of maintenance-food deprivation and, again, found no significant differences in learning or in performance as measured by activity, direction of activity, or rate of responding. There was some evidence, not statistically significant, that the most famished subjects were a bit overeager and that intense drive exerted a mildly inhibitory effect on learning efficiency.

Meyer's data are in complete accord with those presented by Birch, who tested six young chimpanzees after 2, 6, 12, 24, and 48 hr. of food deprivation and found no significant differences in proficiency of performance on six patterned string problems. Observational evidence led Birch to conclude that intense food deprivation adversely affected problem solution because it led the chimpanzee to concentrate on the goal to the relative exclusion of the other factors.

It may be stated unequivocally that, regardless of any relationship that may be found for other animals, there are no data indicating that intensity of drive state and the presumably correlated

amount of drive reduction are positively related to learning efficiency in primates.

In point of fact there is no reason to believe that the rodentological data will prove to differ significantly from those of monkey, chimpanzee, and man. Strassburger has recently demonstrated that differences in food deprivation from 5 hours to 47 hours do not differentially affect the habit strength of the bar-pressing response as measured by subsequent resistance to extinction. Recently, Sheffield and Roby have demonstrated learning in rats in the absence of primary drive reduction. Hungry rats learned to choose a maze path leading to a saccharin solution, a non-nutritive substance, in preference to a path leading to water. No study could better illustrate the predominant role of the external incentive-type stimulus on the learning function. These data suggest that, following the example of the monkey, even the rats are abandoning the sinking ship of reinforcement theory.

The effect of intensity of drive state on learning doubtless varies as we ascend the phyletic scale and certainly varies, probably to the point of almost complete reversal, as we pass from simple to complex problems, a point emphasized some years ago in a theoretical article by Maslow. Intensity of nociceptive stimulation may be positively related to speed of formation of conditioned avoidance responses in the monkey, but the use of intense nociceptive stimulation prevents the monkey from solving any problem of moderate complexity. This fact is consistent with a principle that was formulated and demonstrated experimentally many years ago as the Yerkes-Dodson law. There is, of course, no reference to the Yerkes-Dodson law by any drive-reduction theorist.

We do not mean to imply that drive state and drive-state reduction are unrelated to learning; we wish merely to emphasize that they are relatively unimportant variables. Our primary quarrel with drive-reduction theory is that it tends to focus more and more attention on problems of less and less importance. A strong case can be made for the proposition that the importance of the psychological problems studied during the last fifteen years has decreased as a negatively accelerated function approaching an asymptote of complete indifference. Nothing better illustrates this point than the kinds of apparatus currently used in "learning" research. We have

the single unit T-maze, the straight runway, the double-compartment grill box, and the Skinner box. The single-unit T-maze is an ideal apparatus for studying the visual capacities of a nocturnal animal; the straight runway enables one to measure quantitatively the speed and rate of running from one dead end to another; the double-compartment grill box is without doubt the most efficient torture chamber which is still legal; and the Skinner box enables one to demonstrate discrimination learning in a greater number of trials than is required by any other method. But the apparatus, though inefficient, give rise to data which can be splendidly quantified. The kinds of learning problems which can be efficiently measured in these apparatus represent a challenge only to the decorticate animal. It is a constant source of bewilderment to me that the neo-behaviorists who so frequently belittle physiological psychology should choose apparatus which, in effect, experimentally decorticate their subjects.

The Skinner box is a splendid apparatus for demonstrating that the rate of performance of a learned response is positively related to the period of food deprivation. We have confirmed this for the monkey by studying rate of response on a modified Skinner box following 1, 23, and 47 hr. of food deprivation. Increasing length of food deprivation is clearly and positively related to increased rate of response. This functional relationship between drive states and responses does not hold, as we have already seen, for the monkey's behavior in discrimination learning or in acquisition of any more complex problem. The data, however, like rat data, are in complete accord with Crozier's finding that the acuteness of the radial angle of tropistic movements in the slug *Limax* is positively related to intensity of the photic stimulation. We believe there is generalization in this finding, and we believe the generalization to be that the results from the investigation of simple behavior may be very informative about even simpler behavior but very seldom are they informative about behavior of greater complexity. I do not want to discourage anyone from the pursuit of the psychological Holy Grail by the use of the Skinner box, but as far as I am concerned, there will be no moaning of farewell when we have passed the pressing of the bar.

In the course of human events many psychologists have children, and these children always behave in accord with the theoretical position of their parents. For purposes of scientific objectivity the boys are always referred to as "Johnny" and the girls as "Mary." For some eleven months I have been observing the behavior of Mary X. Perhaps the most striking characteristic of this particular primate has been the power and persistence of her curiosity-investigatory motives. At an early age Mary X demonstrated a positive valence to parental thygmotatic stimulation. My original interpretation of these tactual-thermal erotic responses as indicating parental affection was dissolved by the discovery that when Mary X was held in any position depriving her of visual exploration of the environment, she screamed; when held in a position favorable to visual exploration of the important environment, which did not include the parent, she responded positively. With the parent and position held constant and visual exploration denied by snapping off the electric light, the positive responses changed to negative, and they returned to positive when the light was again restored. This behavior was observed in Mary X, who, like any good Watson child, showed no "innate fear of the dark."

The frustrations of Mary X appeared to be in large part the results of physical inability to achieve curiosity-investigatory goals. In her second month, frustrations resulted from inability to hold up her head indefinitely while lying prone in her crib or on a mat and the consequent loss of visual curiosity goals. Each time she had to lower her head to rest, she cried lustily. At nine weeks attempts to explore (and destroy) objects anterior resulted in wriggling backward away from the lure and elicited violent negative responses. Once she negotiated forward locomotion, exploration set in, in earnest, and, to her parents' frustration, shows no sign of diminishing.

Can anyone seriously believe that the insatiable curiosity-investigatory motivation of the child is a second-order or derived drive conditioned upon hunger or sex or any other internal drive? The S-R theorist and the Freudian psychoanalyst imply that such behaviors are based on primary drives. An informal survey of neo-behaviorists who are also fathers (or mothers) reveals that all have observed the intensity and omnipresence of the curiosity-investiga-

tory motive in their own children. None of them seriously believes that the behavior derives from a second-order drive. After describing their children's behavior, often with a surprising enthusiasm and frequently with the support of photographic records, they trudge off to their laboratories to study, under conditions of solitary confinement, the intellectual processes of rodents. Such attitudes, perfectly in keeping with drive-reduction theory, no doubt account for the fact that there are no experimental or even systematic observational studies of curiosity-investigatory-type external-incentive motives in children.

A key to the real learning theory of any animal species is knowledge of the nature and organization of the unlearned patterns of response. The differences in the intellectual capabilities of cockroach, rat, monkey, chimpanzee, and man are as much a function of the differences in the inherent patterns of response and the differences in the inherent motivational forces as they are a function of sheer learning power. The differences in these inherent patterns of response and in the motivational forces will, I am certain, prove to be differential responsiveness to external stimulus patterns. Furthermore, I am certain that the variables which are of true, as opposed to psychophilosophical, importance are not constant from learning problem to learning problem even for the same animal order, and they are vastly diverse as we pass from one animal order to another.

Convinced that the key to human learning is not the conditioned response but, rather, motivation aroused by external stimuli, the speaker has initiated researches on curiosity-manipulation behavior as related to learning in monkeys. The justification for the use of monkeys is that we have more monkeys than children. Furthermore, the field is so unexplored that a systematic investigation anywhere in the phyletic scale should prove of methodological value. The rhesus monkey is actually a very incurious and nonmanipulative animal compared with the anthropoid apes, which are, in turn, very incurious nonmanipulative animals compared with man. It is certainly more than coincidence that the strength and range of curiosity-manipulative motivation and position within the primate order are closely related.

We have presented three studies which demonstrate that monkeys can and do learn to solve mechanical puzzles when no motivation

is provided other than presence of the puzzle. Furthermore, we have presented data to show that once mastered, the sequence of manipulations involved in solving these puzzles is carried out relatively flawlessly and extremely persistently. We have presented what we believe is incontrovertible evidence against a second-order drive interpretation of this learning.

References

- BIRCH, H. C., The relation of previous experience to insightful problem solving. *J. comp. Psychol.*, 1945, 39, 15-22.
- CANNON, W. B., *The Wisdom of the Body*. New York: W. W. Norton & Company, Inc., 1932.
- CROZIER, W. J., The study of living organisms. In C. Murchison ed., *The Foundations of Experimental Psychology*. Worcester, Mass.: Clark Univer. Press, 1929.
- JENNINGS, H. S., *Behavior of the Lower Organisms*. New York: Columbia Univer. Press, 1906.
- KINSEY, A. C., POMEROY, W. B., & MARTIN, C. E., *Sexual Behavior in the Human Male*. Philadelphia: W. B. Saunders Company, 1948.
- LOEB, J., *Forced Movements, Tropisms and Animal Conduct*. Philadelphia: J. B. Lippincott Company, 1918.
- MASLOW, A. H., A theory of human motivation. *Psychol. Rev.*, 1943, 50, 370-396.
- MEYER, D. R., Food deprivation and discrimination reversal learning by monkeys. *J. exp. Psychol.*, 1951, 41, 10-16.
- RICHTER, C. P., Animal behavior and internal drives. *Quart. Rev. Biol.*, 1927, 2, 307-343.
- SHEFFIELD, F. D., & ROBY, T. B., Reward value of a non-nutrient sweet taste. *J. comp. physiol. Psychol.*, 1950, 43, 471-481.
- STRASSBURGER, R. C., Resistance to extinction of a conditioned operant as related to drive level at reinforcement. *J. exp. Psychol.*, 1950, 40, 473-487.
- THORNDIKE, E. L., *Animal Intelligence*. New York: The Macmillan Company, 1911.
- WATSON, J. B., *Behavior: An Introduction to Comparative Psychology*. New York: Holt, Rinehart and Winston, Inc., 1914.
- YERKES, R. M., & DODSON, J. D., The relation of strength of stimulus to rapidity of habit formation. *J. comp. Neurol. Psychol.*, 1908, 18, 459-482.

Manipulatory Motivation in Children^{*†}

This paper is useful here for two reasons: (1) it is a report of direct research based on Harlow's view of motivation (see pp. 86-90) and (2) it shows how laboratory experimentation with animals can be recast as classroom experimentation with children while utilizing the same theoretical assumptions.

In reading this report the student may find it helpful (1) to make a statement of Terrell's hypothesis and show its relationship to Harlow's position (an exercise in deduction), (2) to note how the experiment was designed to furnish evidence to support the hypothesis, (3) to examine the discussion to see if the drive-reduction theory has been explained away, and (4) to refer to the last paragraph in the introduction to the Harlow article (p. 86) to see if the questions raised about Harlow's manipulation drive have been answered at all in this experiment.

It has been demonstrated that monkeys and kittens can solve simple problems when motivated only by the stimuli of the problem itself. Harlow (1953) has stated that such learning results from an external curiosity-manipulative incentive, and that, contrary to the drive-reduction position, such a motive is not dependent upon

* Reprinted with the permission of the author and the American Psychological Association from the article of the same title, *Journal of Comparative and Physiological Psychology*, 52 (1959), 705-709.

† This investigation was supported in part by a research grant, M-1505, from the National Institute of Mental Health, of the National Institutes of Health, United States Public Health Service, and in part by the University of Colorado Council on Research and Creative Work.

any of the homeostatic drives for its development. Harlow (1953) further decries the absence of experiments involving such incentives in children.

Two previous experiments suggest the operation of a manipulatory incentive in children's learning. Terrell and Kennedy (1957) found that children are able to transpose a larger-than concept when they are engaged in transferring a dried bean from one jar to another following each correct response as well as another group of Ss who receive a piece of candy immediately following each correct response. In what was in many respects a replication of the first experiment, Terrell (1958) found that both the learning and transposition of a size concept was as effective with a group of Ss who were rewarded by only a light flash as with other groups which were candy rewarded. Apparently the manipulation required of the Ss of this study, pushing a button in front of the stimulus of their choice, was sufficiently motivating as to make other incentives unnecessary.

In further support of the speculation that a curiosity-manipulatory motive was operating in the latter study, virtually all Ss indicated in a questionnaire given to them following the experiment that they would "rather do something for the fun of it than to get something for doing it." Although subsequent research by the writer and others (Douvan, 1956; Terrell, Durkin, & Wiesley, *in press*) has suggested that engaging in learning for the sheer enjoyment derived from the activities of learning itself is more characteristic of middle-class than lower-class children, it is believed that discrimination-learning situations may be made such that irrespective of social-class membership, learning can be facilitated by increasing the amount of meaningful manipulative activity required of the Ss in their solution of the task. The present experiment provided a test of this hypothesis. Specifically, this experiment compared the learning and transposition of a group of children who were allowed to engage in certain manual manipulations during the solution of the problem with two other groups of Ss, one of which received only a light flash following a correct response, the other being promised a bag of candy upon completing the task. The task learned was a simple button-pushing response to the larger of two three-dimensional geometric objects.

METHOD

Subjects. The Ss were 192 children, 48 in each of the age categories 5, 6, 8, and 9 years, with an equal number of boys and girls in each group. All Ss were from one school in Boulder, Colorado. There were no 7-year-olds available who had not had previous experience with the task employed in this study. It was also felt that Ss older than 9 years would find the task so simple as to mask treatment effects. Finally, the age range used permitted a study of possible CA differences in the effects of the manipulation condition.

Materials. The apparatus has been described in detail elsewhere (Terrell & Kennedy, 1956). There were three pairs of three-dimensional geometric figures in the shape of cubes, cones, and cylinders. The small member of each stimulus set had a basal area of 4 sq. in., while the large member was 8 sq. in. These stimuli are hereafter referred to as the "training stimulus sets." A third cube with a basal area of 16 sq. in. was used in a transposition test, along with the 8-sq.-in. cube. These stimuli are hereafter referred to as the "test stimuli." The order of presenting the stimuli and the position of the positive, large-size stimulus were randomized alike for each S.

Additional apparatus consisted of a box 16 by 24 by 4 in. containing the batteries and circuits necessary to operate a signal light. Two jacks and two push-button mounts were on top of the box. The stimuli were placed in the jacks on each trial. Locked onto the rear edge of the box was a panel board 10 by 16 by $\frac{1}{4}$ in., which contained the signal light. The circuits were arranged so that a correct response, pushing the button at the base of the large stimulus, caused the light to go on.

Design. There were three experimental groups, all of which were rewarded by a light flash after each correct response. The differential reward conditions of the three groups were as follows: the Ss of Group I were told, "After you choose that one that makes the light go on enough times, I'll give you a little sack of candy." Group II Ss were told, "After you choose the one that makes the light go on enough times, I'll give you a little sack of candy. So, every time you make the light go on, we'll make believe that I take this great, big sack of candy—see it? and reach way down deep inside and take out a piece of make-believe candy—do you like this kind of candy?—which I want you to hold. Do you know what you're going to do with that candy? Well, in your other hand I want you to hold this make-believe sack. Got it? Okay, now put your piece of candy in the sack. Now, after you have enough pieces of make-believe candy in that sack you're holding, you can trade them with me for a sack of real candy. Now don't drop the sack!" The Ss of Group III received only a light flash after each correct response. For convenience, these groups will be referred to as the Candy-Promise, Manipulative, and Control Groups, respectively.

There were two levels of chronological age. Five- and six-year-olds

constituted one age group, while eight- and nine-year-olds were the other group. Within each level of chronological age Ss were randomly assigned to the three reward conditions. This two-dimensional design, reward $\times$ chronological age, was replicated, using a second *E*. Thus, the design for purposes of statistical analysis was a $3 \times 2 \times 2$ factorial arrangement. The first *E* conducted the study in the spring of 1957, while the replication was performed in the spring of 1958. Ninety-six Ss were used in each experiment.

Procedure. The Ss were tested individually. Each child received the following instructions: "This is a game where I want you to choose one of these, and then push the button in front of the one you choose. If you are right, this little light will go on. If you are wrong, the light will not go on. Now remember, the game is to see how quickly you can learn to choose the one that makes the light go on. And there's a way you can tell, each time, which one to choose." The last two sentences were repeated after every tenth trial. Immediately after reaching a criterion of nine consecutive correct responses on the acquisition trials, each S was given a four-trial transposition test. The same differential incentive conditions employed during the acquisition trials were continued during the transposition test.

RESULTS

It was planned that an analysis of variance would be performed on each of the two criterional measures: number of trials to the criterion and number of correct responses on the transposition test. Virtually all Ss responded correctly on the transposition test, mak-

Table 1
MEANS AND STANDARD DEVIATIONS OF TRIALS TO CRITERION,
ORIGINAL DATA

(Each total treatment group, $N = 64$)

Age	Manipulation		Promise		Control		Age Totals	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
<i>Experiment I</i>								
5 & 6	6.57	5.43	10.19	9.04	9.62	9.24	8.79	7.90
8 & 9	7.50	8.72	8.00	7.31	11.94	10.22	9.15	8.72
Treatment totals	7.03	7.07	9.09	8.17	10.78	9.73		
<i>Experiment II</i>								
5 & 6	4.87	7.70	11.75	7.91	5.68	5.26	7.43	6.96
8 & 9	5.81	6.54	10.00	8.13	8.93	7.72	8.25	7.46
Treatment totals	5.34	7.12	10.87	8.02	7.30	6.49		
Grand totals	6.19	7.16	9.98	9.85	9.04	8.54		

Table 2

ANALYSIS OF VARIANCE OF NUMBER OF TRIALS TO CRITERION,
TRANSFORMED DATA

Source	df	MS	F
Rewards	2	.879	4.51*
Age levels	1	.216	
Experimenters	1	.536	2.75
Rewards $\times$ Age	2	.061	
Rewards $\times$ Experimenters	2	.064	
Age $\times$ Experimenters	1	.166	
Rewards $\times$ Age $\times$ Experimenters	2	.077	
Within	180	.195	
Total	191		

* $p < .05$

ing this analysis unnecessary. Apparently the task was simple enough to mask treatment differences once learning occurred. The Bartlett test for homogeneity of variance was significant beyond the .01 level of confidence for the acquisition data, making a transformation of the raw scores advisable. A logarithmic ($\log X + 1$) transformation was employed, which resulted in the homogeneity assumption's being met. Table 1 contains the original acquisition data. The summary of the analysis of variance of the transformed data appears in Table 2. No interaction is significant, and only the main-effects test of reward conditions reaches significance at the .05 level. Comparisons between the three experimental groups were made by use of the t test. The Manipulative Group learned the task significantly more quickly than did the Candy-Promise or Control Groups ($p < .05$ for both comparisons). The Candy-Promise-Control Group difference was not significant. It will be seen in Table 1 that there is a rather noticeable difference in the performance of the Control Groups of Experiments I and II. It is believed that the difference in E characteristics accounts for this fact. The E for Experiment II is a very solicitous person. He does not like to observe a child experience difficulty in an experimental situation. It would be reasonable to expect that he may have unwittingly given the S s of the Control Group some cues which facilitated their learning. There is no known reason to assume that there were sample differences in Ex-

periments I and II which could account for the difference in the Control Groups.

DISCUSSION

The superiority of the Manipulative Group over the Candy-Promise and Control Groups supports the results of research with monkeys by Harlow, Harlow, and Meyer (1950), Harlow and McClearn (1954), Harlow, Blazek, and McClearn (1956), and more recently the experiment by Miles (1958) with kittens. It is true that in the present experiment the Manipulative Group Ss were also motivated by the prospect of exchanging their imaginary sack of candy for a sack of real candy. This feature of the present study does not make the effectiveness of a manipulative incentive less convincing in view of the superiority of the Manipulative Group over the Candy-Promise Ss. Additionally, the fact that the Ss who were merely given a light flash following each correct response (Control Group) learned somewhat more quickly than those who were promised candy after learning (Candy-Promise Group), supports the argument that promising the Manipulative Group a sack of real candy in exchange for the imaginary sack which they had pretended to fill contributes nothing to the motivation of the Manipulative Group.

At least two possible explanations may be advanced to account for the superiority of the Manipulative Ss. It appears reasonable to hypothesize, as Harlow (1953) has done previously, an externally elicited manipulatory motive. The fact that Harlow's Ss were free to manipulate the apparatus spontaneously, while the manipulations of the Ss of the present study were limited by the instructions of the *E*, does not appear to be a serious enough difference as to make necessary the postulation of a different kind of motive. In both cases it appears that the behavior was exteroceptively aroused; in the case of the monkeys and kittens, the characteristics of the apparatus alone appeared responsible for their behavior, whereas in the case of the manipulative-condition Ss of the present study, the apparatus plus the instructions were important. There is evidence to suggest strongly that in both instances something about the learning environment stimulated manipulative activity, which in turn had incentive value for the Ss.

A derived-drive theorist may suggest that postulating such a drive here is unnecessary, that the facilitating effects of the manipulations are derived and therefore do not fit the traditional concept of a drive as being solely an internally elicited energizer and not a director of behavior (Brown, 1953). This conception of a drive seems unnecessarily restrictive, particularly when one considers the complexity of human motivational systems. As Harlow has pointed out (1953), "Motivation results from the activation of 'brain centers' by chemical substances or afferent impulses, and it does not matter whether these hormonal effects or nerve impulses are initiated by exteroceptors or interoceptors" (p. 25). Sheffield and Roby (1950), Sheffield, Wulff, and Backer (1951), Hebb (1955), and McClelland (1953) also argue exteroceptively aroused behavior in some situations. If the incentive effect of the manipulations required of the Ss in the present experiment is derived from some internal drive it would appear that they (the manipulations) would be subject to rather rapid extinction, or at least they would be expected to diminish in strength as the experiment progressed. The opposite effect was observed. Interest of the Manipulative Group heightened considerably more than that of the other groups as learning progressed. The vigor of the responses of reaching for the imaginary candy and placing it in the imaginary sacks increased with the number of trials. Several Ss remarked spontaneously that the imaginary sack was getting "heavy" and pretended to transfer it to the other hand.

The possibility of the superiority of the Manipulative Ss resulting at least in part from the facilitating effects of cognitive activity must be considered. Perhaps the imaginative activity engaged in by these Ss increased the fun and/or meaningfulness of the task, thus presumably increasing motivation. In connection with this suggestion it seems worthwhile to make a comparison of the performance of the Manipulative Ss of the present experiment with the Ss who were permitted to transfer a bean following correct responses in the aforementioned Terrell and Kennedy study (1957). Quite possibly the crucial facilitating factor for the Ss of the latter experiment consisted in imaginary activities which may have accompanied the motor activity of transferring the bean. The contribution of this variable may be studied by comparing the performance of a group of Ss instructed merely to *think* about the manipulations required of the Manipulative Ss of the present study, with the performance of

another group of Ss treated in a manner identical to that of the Manipulative Group of the present experiment. A third group of Ss instructed to think about *and* actually perform these manipulations would be expected to surpass both of the previously mentioned groups if both cognitive and manipulatory activities contribute to the motivation of Ss in this type of learning situation.

Whether it is posited that a manipulatory drive, which the writer favors, a derived drive, or an intrinsic incentive brought about by cognitive activity of the Ss accounts for the superiority of the Manipulative Group Ss in the present study, it is obvious that something about the condition of the Manipulative Group resulted in quicker learning than was characteristic of the other groups. The fact that the experiment was replicated, a different *E* being used in the replication, and that the Manipulative Ss were superior for both *Es* adds to the convincingness of this conclusion. This experiment suggests that interesting prospects exist for the systematic study of externally elicited incentives such as manipulation, exploration, and curiosity in children's learning, something which has heretofore been ignored.

Finally, a matter of some importance concerns the likelihood that the simplicity of the task used in the present experiment masked the treatment differences. It has already been mentioned that this fact probably accounts for the negligible treatment differences in the transposition data. If this is true, one would, of course, expect even greater differences in favor of a manipulative group if more complex discrimination learning tasks were employed.

SUMMARY

This experiment compared the speed of learning and consistency of transposition of a group of kindergarten and elementary school children who are allowed to engage in certain manual manipulations during the solution of a problem with two other groups of Ss, one of whom received only a light flash following a correct response, the other being promised a bag of candy upon completing the task. The task learned was a simple button-pushing response to the larger of two three-dimensional geometric objects.

For the acquisition data, Ss who were allowed to engage in manual manipulations (the Manipulative Group) learned the task

significantly more quickly than did the other two experimental groups. No interaction or age-level difference was significant. Virtually all Ss transposed on every trial, making analyses of these data unnecessary.

It is suggested that either the manual manipulations or the cognitive activity, or both, of the Manipulative Group Ss accounts for their superior performance.

References

- BROWN, J. S., Problems presented by the concept of acquired drives. In *Symposium on Motivation*. Lincoln, Nebraska: Univer. Nebr. Press, 1953.
- DOUVAN, E., Social status and success striving. *J. abnorm. soc. Psychol.*, 1956, 52, 219-223.
- HARLOW, H. F., Mice, monkeys, men, and motives. *Psychol. Rev.*, 1953, 60, 23-32. (a)
- HARLOW, H. F., Motivation as a factor in the acquisition of new responses. In *Symposium on Motivation*. Lincoln, Nebraska: Univer. Nebr. Press, 1953. (b)
- HARLOW, H. F., BLAZEK, N. C., & MCCLEARN, G. E., Manipulatory motivation in the infant rhesus monkey. *J. comp. physiol. Psychol.*, 1956, 49, 444-448.
- HARLOW, H. F., HARLOW, M. K., & MEYER, D. R., Learning motivated by a manipulatory drive. *J. exp. Psychol.*, 1950, 40, 228-234.
- HARLOW, H. F., & MCCLEARN, G. E., Object discrimination learned by monkeys on the basis of manipulation motives. *J. comp. physiol. Psychol.*, 1954, 47, 73-76.
- HEBB, D. O., Drives and the C.N.S. (conceptual nervous system), *Psychol. Rev.*, 1955, 62, 243-254.
- MCCLELLAND, D. C., ATKINSON, J. W., CLARK, R. A., & LOWELL, W. L., *The Achievement Motive*. New York: Appleton-Century-Crofts, 1953.
- MILES, R. C., Learning in kittens with manipulatory, exploratory, and food incentives. *J. comp. physiol. Psychol.*, 1958, 51, 39-42.
- SHEFFIELD, F. D., & ROBY, T. B., Reward value of a non-nutritive sweet taste. *J. comp. physiol. Psychol.*, 1950, 43, 471-481.
- SHEFFIELD, F. D., WULFF, J. J., & BACKER, R., Reward value of copulation without sex drive reduction. *J. comp. physiol. Psychol.*, 1941, 44, 3-8.
- TERRELL, G., JR., The role of incentive in discrimination learning in children. *Child Develpm.*, 1958, 29, 231-236.
- TERRELL, G., JR., DURKIN, K., & WIESLEY, M., Social class and the nature of the incentive in discrimination learning. *J. abnorm. soc. Psychol.*, in press.

TERRELL, G., JR., & KENNEDY, W. A., Discrimination learning and transposition in children as a function of the nature of the reward. *J. exp. Psychol.*, 1957, 53, 257-260.

M. C. WITTROCK

University of California at Los Angeles

Set Applied to Student Teaching*

The concept of *set* is important in psychology and, according to Wittrock, it will probably attract more attention in education than it has up to the present. If one inspects the following numbers

6

4

one knows that one has a choice of 2, 10, and 24 as possible answers. Which response is made depends on one's *set*, whether it is a set for addition, subtraction, or multiplication. Faced with particular situations and problems we are predisposed to respond in particular ways. Set surely influences the choices we make. Duncan has suggested that a problem sometimes remains a problem only as long as we keep the wrong set (p. 220). He has urged more investigations of the possible beneficial effects of appropriate and general sets. A set for "learning how to learn" may be the most useful general set for problem solving (p. 222).

The investigation reported here presents some convincing evidence of the motivational aspects of an appropriate set. Since the subjects in this investigation were student teachers, the reader should find identification easy. The student may try to reconstruct imaginatively how he might have behaved if he were one of the student teachers in the experimental group. How might his motives and efforts have been influenced? Bugelski has suggested that the motivational problem in edu-

* Published with the permission of the author.

cation may finally boil down simply to obtaining students' attention and properly directing it (pp. 83-84). Wittrock's demonstration of the use of set may offer some evidence to support Bugelski's statement.

The following questions may be of interest to the student: (1) How important were grades as sources of motivation for the student teachers in the experimental group? (2) How does the use of grades here compare with that of Bostrom in his experiment on attitude change (pp. 129-135)? (3) What seems to be the relationship between set and conditioning (discussed in the following chapter)? (4) In Chapter Ten the range of possible educational objectives is discussed in connection with Bloom's *Taxonomy* (pp. 569-574). How might this classification be useful if this experiment were repeated?

A set is assumed to increase the probability of the occurrence of certain responses and to decrease the probability of occurrence of other responses usually through selecting, directing, or organizing some part of experience. The various uses of the term "set" range from motor set to goal set. Included between these two extremes are, among others, learning set, response set, task set, and methods of problem solving aroused by directions or by problem situations. For reviews of the concept see Young (1961, pp. 264-278) and Gibson (1941).

There has been much research in psychology on set, but comparatively little attention has been given to it in education. This is surprising since school teachers dispense verbal instructions and give other sets to students almost daily. One method of developing a set in subjects is by appropriately reinforcing trial and error situations (Harlow, 1949) to allow the subjects to "discover" when certain sets are appropriate. Another method of developing sets in subjects is simply to use some form of "reception learning," e.g., verbal instructions (Wittrock, in press). The present study employed the latter technique. As used in this study, *set* referred to the temporary influence upon the behavior of the student teachers, which was produced by certain verbal statements designed to make explicit a particular teaching objective.

In this study reinforcement, i.e., final course grades for the student teachers in the experimental group, was made contingent upon the gain that their secondary school students evidenced on standardized achievement tests given as pretests and posttests.

It is hypothesized that an explicit set to teach for pupil gain on a standardized achievement test results in a change in the teachers' behavior, which produces greater pupil gain than does a comparable procedure that does not specify pupil gain as the criterion.

METHOD

Subjects. The experimental and control groups of 14 student teachers each were divided by subject matter as follows: English—three, American government and history—four, and social studies—seven. There were seven women and seven men in the experimental group. Nine of the experimental group were teaching in their major field, and five of the group were teaching in their minor field. The control group consisted of eight women and six men: 10 were teaching in their major field, and 4 were teaching in their minor field. Five factors were used to match the student teachers individually: teaching major, sex, age, a measure of ability (Cooperative Test Division, 1950), and the public school in which the student teaching was performed. All of the student teachers were regularly enrolled in the University of California, Los Angeles, teacher education program.

A total of 787 junior high and senior high school students were enrolled in the classes taught by the 28 student teachers mentioned above. The 395 students of the experimental group were divided as follows: students in social studies—215, students in American government and history—99, and students in English—81. In the control group there were 392 students of which 199 were enrolled in social studies, 106 in American government and history, and 87 in English. The students in social studies were enrolled in four junior high schools, and the students in American government and history and in English were enrolled in one senior high school. All five of the schools lie in the area commonly referred to as West Los Angeles. None of the above classes were homogeneously grouped according to ability or achievement.

Materials. In the social studies classes the Cooperative Social Studies Test, Grades 7, 8, and 9 (Lloyd, 1948), Parts II and III, were used for both pretests and posttests. In the area of American government, the Cooperative American Government Test (Haefner, 1947) was used for both pretests and posttests. In the American history classes the Cooperative American History Test (Berg, 1948) was used. In all three of the above areas Form X of the appropriate test was used for the pretest, and Form Y of the appropriate test was used as the posttest. In

the English classes the Cooperative English Tests, Parts I and II, which together comprised the Test of English Expression (Cooperative Test Division, 1960), were used. Form 2A was used as the pretest, and Form 2B was used as the posttest. The measure of affectivity toward the subject matter consisted of a five-point scale ranging from one, like very much, to three, do not like or dislike, to five, dislike very much.

Procedure. Three weeks before the beginning of the spring semester, 1961, the college supervisors of the student teachers were informed that an experiment in student teaching and in educational psychology was planned for the next semester. The cooperating, i.e., training, teachers of the student teachers in the experimental group were each told shortly after the beginning of the spring semester that the student teacher was participating in an experiment and that he was to be graded in student teaching on the basis of his pupils' gain. They were told further that a pretest and a posttest would be given during the semester to the secondary school students and that these tests would be objective, standardized tests over the appropriate subject matter. They were not told which standardized tests were to be used. The training teachers of the student teachers in the control group received a letter with the same information as above except that no mention was made of a grading procedure for student teachers. All of the training teachers were informed in these initial letters that they would be given each secondary school student's results of each of the tests and that the experiment would be explained in detail to them at the end of the semester.

At the first meeting of the class in educational psychology, all of the student teachers in the areas of social studies, American government and history, and English were informed that they were to be enrolled in a special educational psychology section and that they were to be part of an experiment on student teaching and educational psychology. All these teachers were required to become part of the experiment. Two days later, on the first day of the special section of educational psychology, the student teachers were told that their final course grades in educational psychology and in student teaching would be determined by the amount of improvement in performance on a standardized achievement test, which they could obtain with their students, as compared with the average amount of improvement of a group of secondary school students taught by student teachers who had been matched to them individually. In order to insure that this set to teach for pupil gain continued to exist in the students, and, of course, to teach educational psychology, the special class continued to meet for 2 hours each week throughout the semester. The "regular" sections of educational psychology also met 2 hours each week. The content of the courses was parallel. The author taught both the regular and the special sections of the course. At no time were the student teachers

coached regarding pupil gain or how to teach for it. The student teachers were not allowed to see which tests had been used or which tests were to be used in their classrooms; they knew only that standardized achievement tests were being used.

During the second week of the semester, letters were sent to each of the training teachers to the effect that a pretest would be given in their classrooms at a day convenient to them sometimes during the fourth week of the semester. The teachers were all told to inform their secondary school students that a standardized achievement test would be given in their classes, that this test was a part of an experiment, and that their scores on these tests would not be used in any way to jeopardize their possibilities of entering college. The students were also told that their scores on these tests would be made available to their teachers. During the fourth week of the semester, the pretest was given to the classes of the experimental and control groups. In each class the experimenter introduced himself as a representative of the University of California and reiterated the fact that the scores would not be used in any way to influence the students' possibilities of entering college but that the scores would be made available to the students' teachers. The experimenter then read aloud the standardized directions from the test manual, and the students were given the time allotted in the standardized directions to complete the examinations. The examinations were all scored by machine according to the directions given in the test manual.

A posttest was given 2 weeks prior to the end of the semester. All training teachers and student teachers were again notified in advance about the examination in a manner identical to the manner indicated above. The procedure that was followed on the test day was identical to the procedure outlined above with one exception. The students were asked to rate on a five-point scale ranging from "like very much" to "dislike very much" their attitudes toward the subject matter which they had studied that semester.

After the experiment was completed all training teachers and college supervisors were informed by letter of the nature and results of the experiment. At the end of the semester the student teachers were graded in educational psychology solely on the basis of pupil gain as compared with the control group teachers' pupil gain. In student teaching, pupil gain was combined with the ratings of the supervising teachers and of the training teachers to obtain the final student teaching grade. In no case did the pupil gain score result in a lowering of any student teacher's final grade.

RESULTS

Table 1 presents the pretest data. The *t* test between the means of the respective experimental and control subgroups indicated that

there was no significant difference in initial performance between them ($p > .05$, two-tailed).

Table 1
A COMPARISON OF THE EXPERIMENTAL AND CONTROL GROUPS'
PRETEST SCORES

Groups	<i>M</i>	<i>SD</i>	<i>t</i>	<i>df</i>
Social studies, experimental	25.11	13.76		
Social studies, control	25.57	13.21	.11	412
Government and history, experimental	25.93	13.13		
Government and history, control	23.39	10.47	1.53	203
English, experimental	46.59	9.29		
English, control	47.76	10.17	-.77	166

The difference scores in Table 2 were computed by subtracting each student's pretest score from his posttest score. Homogeneity of variance tests (Lindquist, 1953, p. 40) for all the comparisons listed

Table 2
DIFFERENCE SCORES BETWEEN PRETEST AND POSTTEST FOR THE
EXPERIMENTAL AND CONTROL GROUPS

Groups	<i>M</i>	<i>SD</i>	<i>t</i>	<i>df</i>
Entire experimental group	7.69	8.03		
Entire control group	5.51	8.33	3.74****	785
Social studies, experimental	10.08	7.48		
Social studies, control	8.69	6.60	2.01*	412
Government and history, experimental	5.79	8.50		
Government and history, control	4.22	8.74	1.31	203
English, experimental	3.68	6.68		
English, control	-0.27	8.20	3.43****	166
Entire experimental ^a	7.49	3.45		
Entire control ^a	5.51	4.43	2.89**	13

^a $N = 14$. See Results section of text.

* Significant at .05 level.

** Significant at .02 level

**** Significant at .001 level.

in Table 2 produced no statistical reason ($p > .05$) to discredit the assumption of homogeneity of variance. From Table 2, the difference between the means of the experimental and control groups was in favor of the experimental group as predicted ($p < .001$, two-tailed). The English experimental and control groups' means differed significantly ($p < .001$, two-tailed) in favor of the experimental group. The social studies experimental and control groups' means differed also ($p < .05$, two-tailed) in favor of the experimental group. The difference between the history and government experimental and control groups' means was not statistically significant at the .05 level. However, the difference was in the predicted direction.

The above t tests between the experimental and the control groups were appropriate to test the hypothesis mentioned in the introduction. However, a more rigorous criterion of whether or not a difference existed between the experimental and control group of teachers was obtained as follows: by using only the mean of the difference scores for all secondary school students in each student teacher's class to measure each student teacher's teaching performance, the experimental group of teachers was compared with the control group of teachers (see Table 2). According to this procedure, the teachers of the experimental group averaged 7.49 points gain between the pretest and the posttest. The control teachers averaged 5.51 points gain between these same two tests. By use of the standard error of the difference formula for individually matched groups, a t of 2.89 ($p < .02$, two-tailed) was found.

The posttest mean scores for the 14 experimental classes and the 14 control classes were compared by analysis of covariance. The posttest mean scores were adjusted for the pretest mean scores. The resulting F was 4.40, $df = 1/25$, $p < .05$.¹

Table 3 presents the analysis of the affectivity ratings of the secondary school students toward the subject matter they studied during the semester. No significant difference was found ($p > .05$, two-

¹ However, from an inspection of graphs of these scores and the difference scores, it is questionable whether or not the experimental group and control group were each normally distributed. From the difference scores, a sign test was computed between the matched pairs in the experimental group and in the control group. Of the 14 matched pairs, 12 of the differences are positive and 2 of the differences are negative. This is statistically significant at the .01 level.

Table 3

AFFECTIVITY TOWARD SUBJECT MATTER FOR THE EXPERIMENTAL
AND CONTROL GROUPS

Groups	<i>M</i>	<i>SD</i>	<i>t</i>	<i>df</i>
Entire experimental group	2.71	1.10	.15	742
Entire control group	2.58	1.13		
Social studies, experimental	2.49	1.14	.21	393
Social studies, control	2.46	1.11		
Government and history, experimental	2.61	1.11	-.07	182
Government and history, control	2.62	1.11		
English, experimental	3.39	1.19	3.13***	163
English, control	2.81	1.17		

Note.—Possible scores ranged from 1 (like very much) to 5 (dislike very much).

*** Significant at .01 level, two-tailed.

tailed) between the entire experimental and entire control groups. However, when subgroup means were compared, the English students in the experimental group scored significantly lower ($p < .01$, two-tailed) than did the English students in the control group.

DISCUSSION

A statistically significant difference between the entire experimental and control groups appeared when the achievement test data were analyzed by any of the statistical procedures mentioned above. The interpretation of the above results must first include a consideration of some limitations that were inherent in the study.

The student teachers in the experimental group were taught educational psychology in a small group situation rather than in a large lecture situation such as was used with the members of the control group. There is, therefore, a possibility that the Hawthorne effect (Roethlisberger & Dickson, 1939) was in operation in this study and that simply giving personal attention to the members of the experimental group may have facilitated their performance. Intact groups were used in this study. However, experimental and control groups each contained samples of each of the same five schools in operation. Further, when one compares the experimental group of

14 student teachers per se with the control group of 14 student teachers, use of intact groups is somewhat more defensible, because each intact group then provided only one measure for the entire group instead of a number of measures equal to the N of the group. Standardized achievement tests were used as the bases for measuring the results of the experiment. The disadvantages of using standardized achievement tests as measures of classroom learning are well known (Thorndike & Hagen, 1961, pp. 310-314). Two of the more important disadvantages of standardized tests are as follows: (a) the test may well lack validity and be insensitive to some important educational process and product outcomes of classroom teaching,² and (b) use of standardized tests as measures of teaching outcomes may imply that the standardized test is a measure of good teaching or of teaching effectiveness.

To measure the effects of the set used in this study, well-constructed standardized achievement tests were appropriate product criteria. The second of the above criticisms does not apply to this study because no attempt is made here to state that improvement on a standardized achievement test is evidence of "good teaching." An evaluation of teaching must include many complex learning processes and learning products. It is suggested here that teaching can be studied separately from the value judgment regarding which directions classroom learning should pursue.

Another limitation is that the secondary school students of the study represented a middle class socioeconomic level. All five schools of this study were located in West Los Angeles.

With the previous limitations in mind, the results of this study are interpreted to provide evidence that the concept of set can be of use in the teaching of social science material to middle class secondary school students.

How this set may have affected behavior presents an interesting problem. The teachers in the experimental group were given some motivation and *direction* not made available to the members of the control group. The direction, namely to teach for pupil gain, was relevant to the criterion measure that was used in this study. The set

² For a discussion of process, product, and presage goals of teacher education, see pages 1482-1486 of the *Encyclopedia of Educational Research* (Mitzel, 1960).

probably contained a cue factor. The cue followed from making explicit the objective to be obtained by the end of the semester. The objective probably conveyed information about the kinds of subject matter which would be sampled on the pretest and on the post-test.

The results of this study are consistent with the operant conditioning paradigm mentioned in the introduction in that the student's behavior was shaped by making reinforcement contingent upon certain responses. Certainly, reinforcement for the student teachers was contingent upon their gaining pupil improvement on a standardized achievement test. The student teachers apparently "shaped" the behavior of their students without developing a dislike for the subject matter of American government, history, and social studies. However, the English teachers in this study, who achieved the highest pupil gain of any of the three subgroups of student teachers of the experimental group, also produced a negative effect upon their students' attitudes toward the subject matter of English. The study implies that by making goals evident and explicit, rather than vague and un verbalized, the behavior of teachers can be changed, but sometimes with undesirable effects upon the attitudes of the students.

References

- BERG, H. D., *Cooperative American History Test*. Princeton: Educational Testing Service, 1948.
- COOPERATIVE TEST DIVISION, *Cooperative English Test, Form S*. Princeton: Educational Testing Service, 1950.
- COOPERATIVE TEST DIVISION, *Cooperative English Test, English expression, Form 2*. Princeton: Educational Testing Service, 1960.
- GIBSON, O. O., A critical review of the concept of set in contemporary experimental psychology. *Psychol. Bull.*, 1941, 38, 781-817.
- HAEFNER, J., *Cooperative American Government Test*. Princeton: Educational Testing Service, 1947.
- HARLOW, H. F., The formation of learning sets. *Psychol. Rev.*, 1949, 56, 51-65.
- LINDQUIST, E. F., *Design and Analysis of Experiments in Psychology and Education*. Boston: Houghton Mifflin, 1953.
- LLOYD, EUNICE, *Cooperative Social Studies Test, for Grades 7, 8, and 9*. Princeton: Educational Testing Service, 1948.

- MITZEL, H. E., Teacher effectiveness. In C. W. Harris (Ed.), *Encyclopedia of Educational Research*. New York: Macmillan, 1960, 1482-1486.
- ROETHLISBERGER, F. J., & DICKSON, W. J., *Management and the Worker*. Cambridge: Harvard Univer. Press, 1939.
- THORNDIKE, R. L., & HAGEN, ELIZABETH, *Measurement and Evaluation in Psychology and Education*. New York: Wiley, 1961.
- WITTROCK, M. C., Set to learn and proactive inhibition. *J. educ. Res.*, in press.
- YOUNG, P. T., *Motivation and Emotion*. New York: Wiley, 1961.

IAN K. WATERHOUSE, *University of Melbourne*

IRVIN L. CHILD, *Yale University*

Frustration and the Quality of Performance*

Ego-involvement has been the subject of much experimental study of motivation and is a motivational source accessible to teacher manipulation. When teachers ask, for example, "How do I get the students involved?" they often mean personally involved. Experimental studies of *frustration* frequently fall into this category, since frustration is viewed as a threat to the ego. It is hardly deniable that teachers and classrooms produce frustration. In fact, some teachers use the amount of frustration as an indicator of success: the more frustrated the students are, the better! Progressive educators, however, have bitterly opposed punitive practices, including frustration, on the general contention that only happy learners are good learners. Even Skinner has warned us about "aversive practices" and has attempted to design learning programs which are easy enough

* R. B. Bugelski, *Psychology of Learning* (New York: Holt, Rinehart and Winston, Inc., 1961), p. 461.

to help the student avoid all mistakes (see pp. 177-179). Bugelski, however, has contended that learning cannot be sufficiently motivated without anxiety and that the teacher's job is not to remove anxiety from the learning situation but rather to regulate its level so that there is neither too much nor too little.*

The following experiments are studies of how and why frustration affects the quality of ongoing performance. The student should note how the type of frustration employed is related to the concepts of reinforcement, discussed in the following chapter. He should also note how complicated the answers to the questions are. Psychological experimentation never gives endorsement to panaceas. There is no prescription here that we should or should not frustrate students in order to improve their learning. The answer is more restricted and complicated. The student may want to ask whether or not teachers should ever resort to deliberate frustration and, if so, with whom and when?

What is the effect of frustration on the quality, or general adequacy, of ongoing performance? At times frustration seems to lead to better, more effective performance; at other times it seems to produce disorganization and less effective performance. Separate demonstration of these two opposite effects is to be found in the experimental literature. Some attention has been given, as by Barker, to a theoretical solution of the problem of how such diverse effects may rise from frustration; but little empirical research has been directed at the problem.

In previous publications, we have suggested, in effect, that this problem may be partly solved by the following proposition: *Frustration will produce a decrease in the quality of ongoing performance, to the extent that the frustration evokes other responses which interfere with that ongoing performance.* We have argued that the results of a well-known experiment by Barker, Dembo, and Lewin, when re-

* Reprinted and abridged with the permission of the authors and publisher from the article of the same title, *Journal of Personality*, 21 (1953), 298-311.

analyzed in the light of this proposition, tend strongly to confirm it. We report here the results of an experiment designed explicitly to test certain implications of this proposition. (We do not mean to imply that the net effect of frustration is necessarily, or even usually, a decrease in quality of performance. Frustration may at the same time, for example, improve performance through increasing motivation. We are simply postulating here that to the extent that frustration leads to interfering responses it will reduce the quality of performance below what it would be in the absence of those interfering responses.)

General Character of the Experiment

The experiment was done with 64 students attending a summer session at a state university; 40 subjects were men and 24 were women, and the sexes were assigned in a constant proportion to the various experimental conditions. Most subjects were close to 20 years of age, though a few were considerably older. The subjects were obtained as volunteers who were to be paid for their services.

The performance which was studied in these subjects was average performance on a variety of tasks which were administered in an individual testing session with each subject, and which were introduced as a battery of tests to measure over-all psychological fitness. For purposes which will be mentioned later, it was desired that some of the tasks should be motor and that some should be very routine intellectual tasks; but, to insure an adequate degree of ego-involvement, one-half of the tasks (6 out of 12) were clearly of the sort that might be found in an intelligence test (reasoning, mental arithmetic, sentence arrangement, picture arrangement, anagrams, and nonverbal analogies). Each of the 12 tasks occupied from 2 to 5 minutes, and provided a quantitative performance score. Since these scores were quite diverse, standard scores were used for each task in order to permit treating the various tasks together. The 12 tasks were given in varying order for different subjects.

The subjects were divided into two main experimental groups, which differed in what the examiner said to them about their performance on these 12 tasks. All the subjects first performed four other, preliminary tasks (not identified to them as being prelimi-

nary); at the end of these preliminary tasks, and then again repeatedly through the course of the 12 crucial tasks, the experimenter made remarks and evaluations to the subject designed to create one of two different effects. For one-half of the subjects (selected at random), the experimenter reported that their performance was very poor and continued at intervals to restate this in various ways; the experimenter was not deliberately insulting in manner, and simply tried, through his evaluations and comments on the subject's performance, to frustrate the subject's desire to feel that he was doing well, or even passably, on the tests. This group will be termed the *Frustration Group*. In an effort to keep the groups alike with respect to orientation toward achievement, evaluative comments were also made to the other half of the subjects; the comments were, however, intended to convey to the subject an impression of mild success, an idea that he was doing just about as well as he might realistically have expected to do. This group will be called the *Neutral Group*, as the aim was, so far as possible, to keep the subjects in an unemotional state, neither frustrated nor elated, while still keeping them oriented toward achievement. The difference in treatment of the frustration group and the neutral group will be referred to as the frustration variable.

From the general hypothesis which guided this study, we predicted that the effect of the frustration variable upon test performance would vary with certain other conditions which were simultaneously varied. These other conditions will be described in turn as the results are presented.

Personality and Effects of Frustration

Our general hypothesis is that frustration will produce a decrease in the quality of ongoing performance, to the extent that the frustration evokes other responses which interfere with that ongoing performance. Among such interfering responses, some of the most important are likely to be internal responses to the fact of frustration—e.g., responses of worry or concern about the frustration. Among the factors influencing the occurrence of such internal responses, a conspicuous factor is likely to consist of stable habits of the individual in response to frustration of any sort. A prediction

was therefore made that the effect of frustration upon quality of performance would vary with the extent to which the individual has general habits of responding to frustration with internal responses of a potentially disruptive sort.

We attempted to verify this prediction by making use of a personality questionnaire in which the subject would report about his habits of response to frustrations of any sort. The questionnaire took the form of a series of 150 statements, for each of which the subject was asked to check on a 6-point scale the extent to which the statement was characteristic of himself. The statements relevant to our present purpose were 90 in number, 15 intended to represent each of the 6 following tendencies in behavior.

- 1) Preoccupation—tendency to be preoccupied with thoughts about previous frustrating or humiliating experiences—e.g., "I tend to brood over my failures."
- 2) Defendance—need to justify or rationalize away one's failures—e.g., "I usually try to explain away my failures."
- 3) Aggression—tendency to react to frustration with extrapunitive aggression—e.g., "When I am angry I am inclined to take it out on other people."
- 4) Pessimism—tendency to react to frustration with feelings of pessimism or dejection—e.g., "Frustration makes me feel very blue."
- 5) Self-aggression—tendency to react to frustration with intrapunitive aggression—e.g., "I am apt to be very critical of myself when I fail."
- 6) Distractibility—tendency to experience continuing interference from distracting circumstances—e.g., "In a tight spot others tend to lean on me, because I keep a perfectly clear head." (These items were phrased in terms of nondistractibility and were scored in the reverse direction.)

From a subject's responses to the 15 items in a single one of these categories, a score could be derived for his tendency to have this particular kind of potentially interfering response to frustration. The 6 scores thus obtained for a single subject could then be added together to yield an over-all measure of his tendency to make potentially interfering responses to frustration. For each experimental group, subjects were then divided into two subgroups according to whether they stood above or below the median of the group in this measure of over-all interfering tendencies. These subgroups will be referred to as the High Interference Group and the Low Interference

Group, and the distinction between those groups in personality characteristics as the *interference variable*.

The effect of frustration upon quality of performance did indeed turn out to vary in the predicted manner with this measure of personality. . . . The results most directly pertinent to our hypothesis are summarized in the right-hand column of the table [omitted]. In the High Interference Group, frustration produced a slight decrement in performance; in the Low Interference Group, frustration produced a large increment in performance. This difference in effect of frustration between the two groups is highly significant; the probability that so large a difference in the predicted direction could have arisen by chance is less than 0.001.

One way of interpreting these results would be to suppose that for both groups the presence of frustration increases the drive to do well; that in the Low Interference Group this increase in drive leads to more effective performance, but that in the High Interference Group the increase in drive is balanced by an increase in interfering responses so that the net effect is that of performance rather similar to what would be found in the absence of frustration. In any event, the prediction made from our hypothesis is clearly and strikingly confirmed. Variations in the tendency to make disruptive responses to frustration, as a general personality variable, have a significant influence upon the effect of frustration on quality of performance. High tendency to make disruptive responses is associated with a relative decrement in quality of performance; low tendency to make disruptive responses is associated with a relative increment in quality of performance.

The results presented in Table I [omitted] may also be looked at from a second point of view, as showing the effect of the frustration variable upon the difference in performance between the High Interference Group and the Low Interference Group. This effect is shown in the last line of the table. In the presence of frustration, performance is poorer in the High Interference than in the Low Interference Group, as might be expected from our hypothesis. In the absence of frustration, however, the difference between the two groups is actually reversed. The High Interference Group now performs considerably better than the Low Interference Group. This fact, while not directly pertinent to the hypothesis we are testing, is

pertinent to any further exploration of the more general significance of the personality variable we are dealing with here. It would appear that the people who, according to their self-rating, tend to react to frustration with interfering responses, tend in a test situation without frustration to produce internal responses which motivate them to do well. Those who are low in tendency to react to frustration with interfering responses and hence are able to work at peak efficiency when motivated by frustration, appear on the other hand to be relatively lacking in motivation under nonfrustrating conditions.

Effect of Instructing the Subject to Introspect

In the previous section we have reported the influence of personality characteristics of individual subjects as determiners of the extent to which they respond to frustration with disrupting internal responses. But we also attempted, in the same study, to influence by experimental manipulation the extent to which subjects would respond to the situation with disrupting internal responses. We did this by instructing half of the subjects to introspect carefully and giving no such instructions to the other half of the subjects. This *introspection variable* was incorporated with the frustration variable in a 2×2 factorial design, making 4 experimental groups.

The precise instructions given to the Introspection Group were as follows:

One other thing—we are also interested in finding out more about just how you as an individual think, and about the way in which you function mentally. So right at the very end of the whole series of tests I will ask you to give a detailed account of the thoughts and feelings and difficulties which you may have had during testing. I want you to pay particular attention right along not only to observing what you are thinking and feeling throughout, but also to fixing these things in your memory, so that you will be able to give an accurate and complete answer afterwards to any specific questions about how you have felt and thought. That means you have to rehearse to yourself quite frequently whatever you will want to mention later.

Subjects in the Introspection Group were then reminded of these instructions several times during the session.

While these instructions constituted quite general instructions to introspect, it seemed possible that they might have a selective effect upon internal responses to frustration. That is, instructions to introspect might be particularly effective where a subject was already inclined to be mulling over his ongoing frustrations, and might thus magnify the disrupting effects of these thoughts. Such an effect could be shown by an interaction of the frustration variable and the introspection variable; that is, the disruptive effects of frustration might be more pronounced when instructions to introspect were given than when such instructions were absent. The results do not confirm this prediction at all. The effect of the frustration variable upon performance is very similar in the Introspection Group and the Non-introspection Group (average increments of 3.8 and 2.9 points, respectively); the direction of this insignificant difference is in fact opposite to the prediction.

The same reasoning might alternatively, however, lead to the prediction of a three-way interaction—that is, the interaction between the interference and frustration variable, reported on in the previous section, might be greater in the presence of the introspection instructions than in their absence. Here too the results do not confirm the prediction. The interaction between the interference and frustration variables may be represented by a value of -3.5 for the Introspection Group and -5.3 for the Nonintrospection Group; this difference, opposite to the predicted direction, does not at all approach significance.

A third possible influence of the introspection instructions turned out to be present to a striking degree. This is its interaction with the interference variable (i.e., personality characteristics) for all subjects, both in the Frustration Group and in the Neutral Group. . . . In the High Interference Group the effect of instructions to introspect was to produce a decrement in performance; in the Low Interference Group the effect of instructions to introspect was to produce an increment in performance. This difference in effect is significant at the 1-per-cent level.

The proper interpretation of this result most probably concerns differences in the internal effect of the introspection instructions in people with differing personality characteristics. Thus, a possible account of these differences might be as follows: The High Inter-

ference Group tends to introspect chiefly about how badly they are doing and about the worrying they are doing in consequence; instructions to introspect then increase their tendency to make these responses, which interfere with effective performance. The Low Interference Group is more task-oriented rather than self-oriented in their introspections, and instructions to introspect lead them to pay even closer attention to the task and its demands and thus to perform more effectively than they otherwise would.

Effect of Kind of Task

It seemed reasonable to us to assume that various tasks or activities might differ in the extent to which they are susceptible to disruption by the simultaneous performance of other responses. If there is a true difference in such susceptibility to disruption, a prediction from our general hypothesis is that tasks should vary in the extent to which frustration will produce a lowering of quality of performance. Tasks which by their nature are easily disrupted by any other concurrent responses should show a maximum lowering of quality of performance in the face of frustration; tasks which are generally impervious to disruption should show a minimum lowering of quality of performance in the face of frustration. Certain kinds of differences among tasks should thus provide another means of verifying the hypothesis that the effect of frustration upon quality of performance is in part a function of disruption through interfering responses.

It is because of this reasoning that we included several different kinds of tests among those given to the subjects. As indicated below, we predicted certain effects of two kinds of difference among our tasks.

1. *Complex Motor Tasks vs. Routine Motor Tasks.* Our battery included two tasks which we thought of as complex motor tasks: a dexterity test (placing steel pins, three at a time, into the holes of a pegboard) and a co-ordination test (repeated successive plunging of a stylus into three holes arranged in the form of a triangle on a sloping board). It also included two tasks which we thought of as routine motor tasks (a tapping test and a dynamometer performance measured as improvement over initial performance at the beginning

of the session). Here it was thought that tremor or sweating in response to frustration would be more disruptive of the complex than of the routine motor tasks. The results did not confirm this prediction, as the average effect of frustration was an increment of 3.5 for the complex motor tasks and an increment of 3.3 for the routine motor tasks, a completely insignificant difference. The reason for this failure of prediction may well be that the frustration was not severe enough for tremor or sweating to have been a significant factor in influencing performance on these tasks, or that the two groups of tasks were not in fact sufficiently different in their requirements.

2. *Complex Intellectual Tasks vs. Routine Intellectual Tasks.* There were six tasks of a relatively complex intellectual character, such tasks as might be included in an intelligence test (reasoning, mental arithmetic, sentence arrangement, picture arrangement, anagrams, and nonverbal analogies). There were two tasks which might be called routine intellectual tasks (rapid addition of pairs of one-digit numbers, and writing the letter *a* as many times as possible within two minutes). Successful performance of the former group appears to require complex series of internal mediating responses, whereas the latter group appears to call for much more automatic succession of responses. Hence it was expected that tasks in the former group would be more susceptible to disruption by such internal responses as worry about failure. From this reasoning came two predictions. The first prediction was that the complex intellectual tasks would be more disrupted by the experience of frustration than would the simple intellectual tasks; there was no acceptable evidence to confirm this prediction, the average effects of frustration being +3.0 and +3.5 for these two types of task. The second prediction was that instructions to introspect would interfere with the complex intellectual tasks more than with the simple intellectual tasks; this prediction was confirmed, with results significant at the 1-per-cent level. . . .

Our predictions about the variation of effects with the type of task were thus only in small part confirmed. We are inclined to feel that the predictions may well have some considerable validity none the less, and that we have demonstrated only that they do not hold true for such simple conditions. Where an intellectual task becomes

even more complex and more demanding of close attention to a long series of internal responses, it may well be that the interfering effects of frustration and irrelevant introspective tendencies will become even more pronounced.

Summary of Main Experiment

This experiment was designed to check several possible implications of the hypothesis that lowered quality of performance, as an effect of frustration, results from responses which interfere with ongoing performance of high quality. In one instance, an implication drawn from the hypothesis was fully and accurately confirmed: lowered quality of performance after frustration is clearly associated with the self-reported personality characteristic of making potentially disruptive responses to frustration in general. For two other sets of implications, there was only partial confirmation. Instructions that the subject is to introspect do interact with self-reported personality characteristics in a way consistent with the hypothesis, but fail to interact significantly with the frustration variable or with these two variables together. Predictions that the effect of frustration would vary with the type of task (because of differing susceptibility to interference) were not confirmed for the particular tasks used in the experiment; on the other hand, the effect of instructions to introspect did vary, as predicted from the hypothesis, according to type of task.

Conclusions

We have proposed that frustration will produce a decrease in the quality of ongoing performance, to the extent that the frustration evokes other responses which interfere with that ongoing performance. From this hypothesis we have made several predictions. One prediction, that the effect of the frustration on quality of performance would vary according to a person's general habits of response to frustration, was confirmed both in our main experiment and in a supplementary experiment. This finding is also consistent with the outcome of a study by Mandler and Sarason, in which a different task and a different personality measure were used. It is also con-

sistent with a general interpretation of the effect of frustration upon quality of performance which we based on a critical review of earlier studies.

The other two sets of predictions, one having to do with the effect of instructions to introspect and one having to do with the nature of the task, were only in part confirmed. The negative findings here certainly set limits on the generality with which the hypothesis we have formulated can be applied. Because these predictions had to do with a novel technique (in the case of using instructions to introspect as a means of introducing interference) or with otherwise untested assumptions (in the case of the assumed differences among the four types of task), it is not possible at present to judge satisfactorily how narrow those limits must be.

References

- ALPER, T. G., Memory for completed and incompleted tasks as a function of personality: correlation between experimental and personality data. *J. Personal.*, 1948, 17, 104-137.
- BARKER, R. G., Frustration as an experimental problem. V. The effect of frustration upon cognitive ability. *Character and Pers.*, 1938, 7, 145-50.
- BARKER, R. G., DEMBO, TAMARA, AND LEWIN, K., Frustration and regression: an experiment with young children. *Univ. Ia., Stud. Child Welf.*, 1941, 18, 1-314.
- CHILD, I. L., AND WATERHOUSE, I. K., Frustration and the quality of performance: I. A critique of the Barker, Dembo and Lewin experiment. *Psychol. Rev.*, 1952, 59, 351-362.
- CHILD, I. L., AND WATERHOUSE, I. K., Frustration and the quality of performance: II. A theoretical statement. *Psychol. Rev.* (in press).
- MANDLER, G., AND SARASON, S. B., A study of anxiety and learning. *J. abn. soc. Psychol.*, 1952, 47, 166-173.

ROBERT N. BOSTROM
Western Illinois University

Grades as Reinforcers in the Production of Attitude Change^{*}

This research report describes how rewards, in this case grades, become important sources of motivation. It offers further evidence that internal needs alone do not explain classroom behavior, unless, as in this case, we assume a grade-hunger.

The question of what grades are for, of how they should or should not be used, is a vexing one in the educational world. Should grades be given for achievement only—or on the basis of effort, ability, improvement, etc.? Should grades be given at all if they tend to distract students from the real educational objectives, if poor grades tend to discourage further attempts at learning, if good grades tend to cause future laziness on the part of the student? Should students grade themselves, so that they can learn to appraise their own achievements when teachers are not around? The student can begin here his consideration of a question which will be discussed in greater detail in Chapter 10 (see especially Adkins, pp. 575–586).

The following experiment indicates that grades can be an integral part of the learning experience. The *grades themselves* can be important sources of motivation for the student and result in significant changes in his behavior. The student can refer to this experiment after he has read the discussion of reinforcement theory in the following chapter to see how the hypothesis here was logically deduced, as discussed by McDonald (see pp. 59–62).

^{*} Revision of an article by the author, John Vlandis, and Milton Rosenbaum, entitled "Grades as Reinforcing Contingencies and Attitude Change," *Journal of Educational Psychology*, LII (1961), 112–115. Reprinted with the permission of the author and the American Psychological Association.

The use of grades as motivators in educational practice has received considerable attention in the research literature, but little information is available about the effects of grades on the students who receive them. The present study was designed to examine the effect differential assignment of grades might have on attitudes when the grades were assigned to essays written that expressed attitudes.

The basic hypothesis guiding the study was that grades will serve to affect behavior in the form of a reinforcer. More specifically, a "good" grade should produce repetition of the responses which it follows, while a "poor" grade should reduce the frequency of the preceding responses. In particular, it was predicted that students who received an "A" for a written composition should show attitude change in the direction of the position adopted in the essay, and if the student would receive a "D" the attitude change would be in the opposite direction.

In applying the hypothesis of the reinforcing effects of grades to attitude changes in this context, the formulation presented by Doob is being followed. Doob views an attitude as an implicit, anticipatory response which mediates overt behavior but which in turn is derived from the reinforcement of overt behavior. For Doob, reward or avoidance of punishment may constitute the reinforcing contingency. Accordingly, the reinforcement of attitude-related, overt statements may be expected to be functionally related to changes in measured attitude. Evidence supporting this view has been supplied by Scott. The present study attempts to extend Doob's formulation into the area of educational practices.

Scott has proposed the possibility that verbalization of a position opposed to initial attitude in and of itself may produce change with reinforcement or nonreinforcement leading to stability or extinction of the new response. Janis and King showed that verbalization alone in contrast to the effect of the consequent contingencies could not be examined by Scott because of the absence of a control group which experiences no consequences of their verbalizations. In the present experiment, a group was included that received no grade following their essays.

METHOD

A 40-item questionnaire containing four 10-item attitude scales were administered to 228 students enrolled in Communication Skills classes at the State University of Iowa. Subjects responded to each item on a five-point continuum ranging from strongly agree to strongly disagree. The four attitude scales dealt with federal aid to education, legalized gambling, capital punishment, and socialized medicine. Responses to the scale on federal aid to education were extremely skewed and the capital punishment scale proved to be unreliable. These two scales were discarded from further consideration and served primarily for filler items. The initial administration of the questionnaire was conducted by the classroom instructors.

During a class period, approximately 6 weeks following administration of the scales, the subjects were asked by the experimenter to write essays on particular assigned topics. The subject was instructed by directions appearing at the top of a sheet to write a brief essay supporting a position on either legalized gambling or socialized medicine. Scores on the attitude scales determined the position that was assigned. In each case, the subject was instructed to write supporting the position opposite to that indicated by his pretest scale score, i.e., if the subject's scale score indicated favorability to legalized gambling, he was asked to write in opposition to legalized gambling. The designation of topic to a particular subject was based on the strength of his initial position on the scales. The scale chosen was the one on which the subject had assumed the strongest position relative to the other scales. One-half hour was permitted for the essay. On completion of the essays, the experimenter promised to return grades on the following day.

On a random basis, grades were assigned to the essays. One third of the subject's writing on each topic received a grade of A, one third received a grade of D, and one third was given no grade. This last group was told when papers were returned that the ratings were not completed due to insufficient time. Immediately after returning the essays and grades, the attitude questionnaire was re-administered. Finally, subjects were asked to indicate their satisfaction with the essays. The total number of subjects participating in all phases

of the study numbered 127, of whom 58 wrote an essay on legalized gambling and 69 wrote on socialized medicine.

Scores on each 10-item attitude scale were computed by summing the responses to each item for which a weight from 1 to 5 was given. The total minimum possible score for each scale was 10 and the maximum possible score was 50. The dependent change measure was derived by subtracting each subject's score on the posttest from his score on the pretest. A change in the direction of the position taken on the essay was given a positive sign. Change in the opposite direction was negative. To eliminate negative numbers a constant of 20 was added to all change scores.

RESULTS

It was predicted that subjects who were awarded an A would change on the average in the direction of their essays to a greater extent than subjects who were given a D. Table 1 presents the data relevant to this prediction. Subjects receiving an A changed an average of 31.76 points in the direction of their essays while subjects who were given D changed 25.85 points. This difference is significant at beyond the .01 level. Comparisons of the groups that received a grade with the group that did not, indicates significantly greater change ($p < .05$) for the subjects who received an A than for those given no grade, while no difference is suggested between the subjects obtaining a D and subjects receiving no grade.

Analysis of mean change for each of the issues suggests similar results to those obtained in the over-all analysis although the differences obtained were more striking for the socialized medicine issue than for legalized gambling. An analysis of mean change in relation to initial position indicates that those who had initially assumed a favorable position on each of the issues (i.e., in favor of legalized gambling or socialized medicine) changed significantly more ($t = 4.73$, $p < .01$) than those who were unfavorable. Nevertheless, the effect of grades on attitude change remains similar to that obtained when direction of initial position is not considered.

Subjects, in addition to responding to the attitude scale related to the topic on which they had written an essay, also responded to the scale for which a related essay had not been written. By com-

paring the change scores of subjects who had written an essay on a particular topic with those who had not written on that topic it is possible to evaluate the effect of essay writing independent of the effect grades.¹ The mean change obtained for subjects who had written a relevant essay was 28.24 in contrast to 22.61 for subjects who had written on the other topic. This difference is significant

Table 1
MEAN ATTITUDE CHANGE

Change ^a	Grade A (N = 42)	Grade D (N = 47)	No Grade (N = 38)
Mean	31.76	25.85	27.11
SD	11.16	7.84	8.03
Differences between groups ^b			
	<i>t</i>	<i>p</i>	
A vs. D:	2.84	< .01	
A vs. No:	2.13	< .05	
D vs. No:	0.73	> .10	

^a A constant of 20 was added to all change scores.

^b Because of heterogeneity of variance, *t* tests were computed employing the procedure recommended by Edwards.

($t = 3.27$, $p < .01$) suggesting that the writing of an essay, independent of grade received, produced change in attitude.

DISCUSSION

The results suggest support for the hypothesis that a "good" grade serves to reinforce the behavior for which it has been administered. Verbalization without a consequent contingency seems to lead to responses similar to those obtained when verbalization is

¹ Comparison of the scores of each of the experimental groups individually with this nonessay writing "control" group would seem to be precluded because of potential unknown effects of simply receiving a grade. Collapsing across groups regardless of experimental condition potentially randomizes such effects.

followed by a "bad" grade. "Cognitive contact" with the opposing side in and of itself does not appear to produce effects approaching those obtained when reward is forthcoming.

Some insight into additional effects of the grades is provided by the results associated with a question as to the satisfaction the subjects experienced with their essays. Of the 52 subjects who received an A, 42 indicated satisfaction with their essays. Twelve of the 57 subjects who received a D were satisfied with their essays. Of the subjects who received no grade, 12 out of 57 were satisfied. A chi square of 41.99 ($df = 2$) indicated that this distribution is significant beyond the .001 level. It is apparent that no difference in satisfaction with their essays appears for those who received a poor grade and those who were given no grade. The no grade condition was an unusual one in class practice and seemed to operate for the subjects as a poor grade, perhaps due to the frustration of failure for the expectancy of receiving some grade to be fulfilled.

Some evidence is provided that verbalization in an incongruent essay is effective in producing attitude change, independent of consequences. This supports the findings of Janis and King on the relation of role playing and attitude change.

The experiment presents evidence only on the effect of grades on one aspect of a complex set of behaviors appearing in the essay writing situation. A qualified generalization can be offered that the administered grades affected in a similar manner many other unmeasured aspects of performance in the situation, i.e., compositional skills, ideational patterns, affect concerning essay writing, etc. Further research is necessary to delineate those behaviors affected in grading situations. Educational practice will profit from understanding of the functional role of grades for other than description of assessment procedures.

SUMMARY

The potential effect of grades as a reinforcing contingency was examined. University students wrote essays defending positions on attitude related issues contrary to their previously assessed positions. Good and poor grades were randomly assigned to the essays and reported to the students. The effect of these procedures on attitude

change were evaluated and good grades were demonstrated to serve a reinforcing role in contrast to the effects of a poor grade or no grade.

References

- DOOB, L. W., The Behavior of Attitudes, *Psychol. Rev.*, 1947, **54**, 135-156.
- EDWARDS, A. L., *Experimental Design in Psychological Research*. (Rev. ed.) New York: Holt, Rinehart and Winston, Inc., 1960.
- JANIS, I. L., and KING, B. T., The Influence of Role Playing on Opinion Change. *J. Abnorm. Soc. Psychol.*, 1954, **49**, 211-218.
- ODELL, C. W., Marks and Marking Systems, in W. S. Monroe (Ed.), *Encyclopedia of Educational Research*. (Rev. ed.) New York: The Macmillan Company, 1950, pp. 711-717.
- SCOTT, W. A., Attitude Change Through Reward of Verbal Behavior. *J. Abnorm. Soc. Psychol.*, 1957, **55**, 72-75.
- SCOTT, W. A., Attitude Change by Response Reinforcement: Replication and Extension, *Sociometry*, 1959, **22**, 328-335. (a).
- SCOTT, W. A., Cognitive Consistency, Response Reinforcement, and Attitude Change. *Sociometry*, 1959, **22**, 219-229. (b).

[CHAPTER 3]

Programed Learning: The Ordering of Learning Situations

Introduction

Of all the recent developments in education, few are likely to have an impact on classroom practice that exceeds that of teaching machines and programed learning. Research findings indicate that teaching-machine programs are as effective as conventional methods of instruction. They seem to do as well in teaching information, and they often do the job in considerably less time. How they will compare in teaching more complex intellectual abilities and skills remains to be seen.

Programing requires of the programmer a thorough mastery of subject matter, both information and organization. It forces him to make decisions about what is essential in a field of knowledge and it demands that he understand the organizing principles in that field. Although the content specialists from the universities and colleges may be building the first programs, it may later appear desirable for teachers to build their own programs as local needs dictate or at least to modify the standard programs. It may very well be that the training of teachers in the years ahead will require courses in programing.

Programing may encourage educators to re-examine the total curriculum. In order to arrange programs or to fit them into a series

of learning experiences it will be necessary to know quite explicitly what our educational objectives are (see pp. 569-575). It may even prompt more modesty and specificity in the statement of educational objectives than in the past. Also, it may help educators to see the curriculum as a sequence of learning tasks which need careful articulation. Furthermore, educational tasks are frequently complex and require analysis into component tasks. Such detailed redesigning of the curriculum may be the chief benefit of programed learning.

Although there may be a renewed concern with subject matter, as the student will see in the subsequent readings, the researchers are less interested in the use of the machine to promote the memorization of materials than in translating subject matter into responses which the student can make and evaluate immediately. Because attention here is on both content and behavior (or learning), the teacher is forced to think simultaneously about the what, how, and why of teaching and learning. As a technological advance, teaching machines may prove to be as great an emancipation of the teacher as the steam engine and electricity were of the worker. Furthermore, although the teaching machine may be described as impersonal, unfeeling, and cold, as teachers supposedly never are, it is the machine, paradoxically enough, which offers one of the best hopes of accommodating individual differences in the wide range of ability, levels of development, etc., which we observe among students in any school classroom. By freeing the student to study independently and at his own rate we have attained a mass tutorial system—the machines replacing the tutors. Yet, the conditions of their most effective use remain to be discovered by the proper experimental investigations.

One still-debated issue in the use of teaching machines is whether Skinner's linear programs are better than Crowder's branching (or intrinsic) programs. In defense of his position, Crowder refers to the greater accommodation of individual differences and the greater amount of potential motivation and responsiveness of his programs when compared with Skinner's (see pp. 164-182). His defense is essentially pragmatic. If the branching program does the job more effectively, why worry about the underlying learning principles of either type of program—especially when the principles have been derived from experimentation with pigeons? Skinner has insisted

that the program should be linear. The defense of his position is based on the concept of reinforcement. Since the topics of operant conditioning and reinforcement are receiving widespread attention in educational and psychological literature, the reader should be clear as to what these concepts are.

The teacher smiles her approval when Sally supplies the phrase that will neatly express her idea; John's selection as valedictorian recognizes his superior academic ability; the students roar their approval when Gus crosses the goal line for the winning touchdown. In all these situations the process of reinforcement is in operation. The smile, the academic recognition, and the roar of approval are the reinforcers or the rewards (in Skinner's usage) for desirable behavior. The reinforcement or reward of act consequent to the act increases the likelihood that the act will be repeated in the future. We would, therefore, expect further *mots justes* from Sally, high academic achievement from John, and touchdowns from Gus.

Skinner has been able to demonstrate the process of reinforcement by training a pigeon to peck at a spot on the wall of the experimental box in which it is enclosed. Skinner describes the process as follows:

We first give the bird food when it turns slightly in the direction of the spot from any part of the cage. This increases the frequency of such behavior. We then withhold reinforcement until slight movement is made toward the spot. This again alters the general distribution of behavior without producing a new unit. We continue by reinforcing positions successively closer to the spot, then by reinforcing only when the head is moved slightly forward and finally only when the beak makes contact with the spot. We may reach this final response in a remarkably short time. A hungry bird, well adapted to the situation and the food tray, can usually be brought to respond in this way in two or three minutes.*

In the case of the pigeon which pecks a designated spot, any particular instance of pecking in the pigeon's life, is called a *response*. But this pigeon has pecked more than once in its lifetime and it has pecked at more things than spotted walls. Such "pecking," regardless of specific occurrences in its lifetime, is called an *operant*. It is the way the pigeon has of "operating" on its environment in order

* Reprinted with permission of the publisher from *Science and Human Behavior* by B. F. Skinner, p. 92. Copyright 1953 by The Macmillan Company.

to get the results it seeks. When Skinner conditions a pigeon, he tries to get it to do more pecking than it would otherwise do. The term, *operant conditioning*, designates simply a change in frequency in particular behavior. When the pigeon pecks on the spot on the wall it is given food, the *reinforcer*. The whole process of rewarding the pigeon with food when it pecks at the spot on the wall is called *reinforcement*. Skinner gives further examples:

While we are awake, we act upon the environment constantly, and many of the consequences of our action are reinforcing. Through operant conditioning the environment builds the basic repertoire with which we keep our balance, walk, play games, handle instruments and tools, talk, write, sail a boat, drive a car, or fly a plane. A change in the environment—a new car, a new friend, a new field of interest, a new job, a new location—may find us unprepared, but our behavior usually adjusts quickly as we acquire new responses.*

The concept of reinforcement is related to several readings in other chapters as well as those here. It is, for example, closely related to motivation. Whether reinforcement or *reward* is viewed as a satisfaction of a need or a reduction of a drive or not, most current theories of motivation require a concept of reward. Also, as indicated in the preceding chapter, what was once a reward, such as a grade of "A" on a term paper, can later become a source of motivation (see pp. 82–83). In this way, motivation is not separable from reinforcement. Its separate treatment in this chapter is justified by its wide application in teaching machines and programed learning, described in the succeeding articles. In this book reinforcement was first mentioned in connection with the selection from *Walden Two*, where a society is created by use of positive reinforcement and where punishment and force (as aversive stimuli) are absent (see pp. 39–40). Gagné and Bolles refer to reinforcement as a *readiness factor* and discuss it in terms of the learner obtaining "knowledge of results" immediately (see pp. 10–20). Rewards as a source of motivation were illustrated by the experiment in which grades were used to obtain attitude changes in students (see pp. 129–135). With regard to the use of the mass media in education, the question of how reinforcement will be provided must be raised.

* Reprinted with permission of the publisher from *Science and Human Behavior* by B. F. Skinner, p. 66. Copyright 1953 by The Macmillan Company.

The Relationship of Readings in Chapter 3

The articles by Verplanck and Lewis are examples of the experimental study of reinforcement as used in the control of human behavior. Lewis' article also introduces the idea of schedules of reinforcement. The remaining articles concern teaching machines and programed learning. Skinner shows how the development of teaching machines is related to the science of behavior and the concept of reinforcement. His article also contains an example of linear programming. Crowder defends a different type of programming—intrinsic or branched. Both Crowder and Gilbert believe that previous laboratory studies of learning can lend little help in the development of programs and programming. Crowder views the teaching program as communication with students. He stresses the need for flexible programming to accommodate individual differences in students. Gilbert describes a quite unconventional type of laboratory research which he believes will yield more information about human learning and proper programming. Ferster and Sapon, using various principles of learning, devised a program for teaching German. They describe the program and report the results of their research.

WILLIAM S. VERPLANCK
Stanford University

The Control of the Content of Conversation: Reinforcement of Statements of Opinion *

Several experimental studies of operant conditioning and reinforcement in human behavior are described in this article. The author, expressly interested in the possibilities of extending

* Reprinted with the permission of the author and the American Psychological Association from the article of the same title, *Journal of Abnormal and Social Psychology*, 51 (1955), 668-676. Footnotes are omitted.

the study of behavior from the laboratory to ordinary life situations, reports the results of the subtle rewarding of certain responses which may occur during a casual conversation. Such experimentation, of course, has vital import for the educational psychologist, who cannot have the usual laboratory conditions to work in. Instead, he must thread his way through school corridors, desks, students, teachers, and principals. And, a factor not to be overlooked, his welcome in the school is often in inverse relationship to the frequency of his visits.

In the discussion of his experiment Verplanck points out that the procedure for operant conditioning may not be as simple and direct as the introduction to this chapter may lead the student to believe. If it is six-year-old Johnny whom we want to condition, we must remind ourselves that Johnny is not an infinitely plastic organism and that he cannot produce any response we may desire. Furthermore, we frequently have only vague notions of what heretofore indifferent aspects of Johnny's environment might suddenly become interesting to and rewarding for him.

Verplanck's experiments in getting individuals to form opinions and make statements on preselected topics are useful illustrations of how teachers can promote particular behavior through operant conditioning. The education student may wish to think about promoting behavior in the classroom through the process of operant conditioning. To determine the conditions in which he can accomplish this establishes a rudimentary learning experiment. In this way the student can begin to see that what can be trusted to science in teaching should not be left to art: Reinforcement, as an unavoidable aspect of all learning situations, should be controlled, to make sure it works for rather than against promoting efficient and appropriate learning.

Some kinds of human behavior have seemed to be resistant to experimental investigation because of both their complexity and their apparent variability. One such class includes the commonplace ac-

tivities of people—for example, whatever the reader was doing just before he picked up this journal. Perhaps talking to someone.

This paper describes the successful experimental application of some principles of operant conditioning in this area; specifically to conversation between two people. The experimental procedure is based on two assumptions. (a) Apparently heterogeneous human verbal behavior falls into comparatively simple operant response classes; hence, any one is susceptible to conditioning. The class of verbal behavior chosen is the *stating of opinions*. (b) Classes of environmental events can be isolated that have the property of altering any behavior on which their occurrence has depended, i.e., some events are reinforcing stimuli. Specifically, under our conditions, statements of *agreement* or *paraphrase* are hypothesized to be reinforcing stimuli for the verbal behavior of a speaker. According to these assumptions, if someone agrees with every opinion of a speaker, the speaker should show a sharp increase in his rate of stating opinions. The *stating of opinions* has been conditioned.

Since it is both interesting and important to obtain changes in behavior that correspond to those termed conditioning when the subject is not aware that he is "being conditioned" (or, indeed, that his behavior is being manipulated in any way) the present experiments were conducted under conditions in which the occurrence of such "insight" was extremely unlikely.

METHOD

General Plan of the Experiment. The experiment was carried out in a series of ordinary conversations between two people, the subject (S) who was not informed in any way that he was taking part in an experiment, and the experimenter (E). The conversations lasted at least a half-hour which was divided into three 10-minute periods.

During the first 10-minute period, once the conversation was under way, E did not reinforce any statement made by S, but determined his operant level of "stating opinions" by ticking off the total number of statements and the number of opinion-statements made by S in successive one-minute intervals. This treatment for the first 10-minute period is labeled O in the first column of Table 1.

In the second 10 minutes, every opinion-statement S made was recorded by E and reinforced. For two groups, E agreed with every

opinion-statement by saying: "Yes, you're right," "That's so," or the like, or by nodding and smiling affirmation if he could not interrupt. This treatment is labeled A, for agreement, in the second column of Table 1. For two other groups, *E* reinforced by repeating back to *S* in paraphrase each opinion-statement that *S* made (labeled P in the second column of Table 1).

Table 1
TREATMENTS FOLLOWED BY EXPERIMENTERS

<i>N</i>	First 10 Minutes		Second 10 Minutes	Third 10 Minutes
5	O—Measure level	operant	A—Reinforce each opinion-statement by agreement	D—Extinguish by disagreeing with each opinion-statement
2	O—Measure level	operant	A—Reinforce each opinion-statement by agreement	E—Extinguish by failing to respond to any statement of <i>S</i> (silence)
6	O—Measure level	operant	P—Reinforce each opinion-statement by paraphrase	D—Extinguish by disagreeing with each opinion-statement
4	O—Measure level	operant	P—Reinforce each opinion-statement by paraphrase	E—Extinguish by failing to respond to any statement of <i>S</i> (silence)
7	A ₁ —Reinforce each opinion-statement by agreement	each	E—Extinguish by failing to respond to any statement of <i>S</i> (silence)	A ₂ —Reinforce each opinion-statement by agreement

In the third 10-minute period, the *Es* attempted to extinguish the opinion-statements of two groups by withdrawing *all* reinforcement, that is, by failing to respond (labeled E for extinguish in the third column of Table 1) in any way to *S*'s speech, and of two other groups by disagreeing with each opinion stated (labeled D in the third column of Table 1).

The design of the experiment is depicted in Table 1. Of the four O-groups of the first period, two become groups in which reinforcement came by agreement (A-groups) in the second period, and two became groups in which reinforcement came by paraphrase (P-groups). In the third period, one of the A-groups was extinguished

by disagreement (D-group) and one by *E*'s silence (E-group). A similar division was made for the P-groups. Thus, each of the four groups can be designated by the combination of treatments provided in the three consecutive periods of conversation.

In a fifth, control group (A_1EA_2), run to insure that any changes in *S*'s rate of stating opinions could not be attributed simply to the passage of time during the experiment, *E* reinforced by agreement *S*'s opinion-statements in the first and third 10-minute periods, and withdrew all reinforcement during the second period.

During the first (O) period for the first four groups, and the E period for the fifth group, *E* asked a "neutral" question ("What did you say?") if *S*'s rate of speaking showed signs of declining. Few such were necessary.

Experimental Situation. The *Es* performed the experiment when and where they could, restricted by only three criteria: (a) that only two persons be present, (b) that there be a clock, and the paper and pencil required for recording, and (c) that enough time be available to both *S* and *E* for them to talk for at least a half hour. The *Es* did not suggest to *Ss* at any time that an experiment was being carried on, and in the rare cases in which an *S* showed signs of suspicion that this was not an ordinary conversation the experiment was terminated (although the conversation was carried on).

Seventeen *Ss* were run in student living quarters, two in restaurants, two in private homes, and one each in a hospital ward, in a public lounge, and over the telephone. In one experiment, contrary to instructions, a third (but uninformed) person was present.

The topics of conversation ranged from the trivial to the "intellectual" and included dates, vacations, Marxism, theory of music, man's need for religion, architecture, Liberace.

Experimenters. Seventeen members of a course in the Psychology of Learning served as *Es*. Twelve were Harvard undergraduates, two were Radcliffe undergraduates, and three (two women and one man) were students in the Graduate School of Education. All the experimenters had had extensive experience in the techniques of conditioning bar-pressing in the rat, and of conditioning chin-tapping in the human. Of the 17 students who undertook the experiment, all were able to collect one or two sets of data as the design demanded.

Subjects and Experimental Groups. Of the 20 men and four women who served as *Ss*, 13 were described by the *Es* as friends,

seven as roommates, one a date, one an uncle, and one a total stranger. In all but four conversations, *S* and *E* were of the same sex. All but six *Ss* were of college age; of these six, four were in the thirties, and two were 55 and 60, respectively.

These *Ss* were distributed over the four experimental groups as follows: OAD, 5; OPD, 6; OPE, 2; OAE, 4; and A_1EA_2 , 7.

There were 20 students in the class, and the design called for *Ns* of 5 and 10, but 3 students reported that they were unable to undertake the experiment, and of the 17 *Es*, one placed himself in the wrong group.

The Response Conditioned. The response selected for reinforcement was the uttering by *S* of a statement or "sentence" beginning: "I think . . .," "I believe . . .," "It seems to me," "I feel," and the like. The *Es* were instructed to be conservative in classifying a statement as an opinion, and to do so only if one or another such qualifying phrase began the statement. (*Es* were aware that the experiment was designed to investigate *Ss'* behavior, and not their own.) No attempt was made to define what constituted a statement or a "sentence" except that *E* should not expect grammatical sentences. These instructions proved adequate; no *E* had difficulty in counting such units of verbal behavior, although doubtless many speech units counted would not parse.

Reinforcing Stimuli. Two classes of reinforcing stimuli were used by the *Es*. The first was *agreement* (A), defined as the experimenter saying "You're right," "I agree," "That's so," or the like, nodding the head, smiling (where *E* did not want to interrupt). The second was repeating back to *S* in paraphrase (P) what he had just said. No further attempt was made to specify paraphrasing. *Extinction* was carried out in one of two ways. In some groups *E* simply refrained from responding in any way to a statement by *S* (E) and in others, he disagreed (D) with each opinion-statement.

The *Es* did not speak, except to reinforce, to disagree, or to "prime" *S* with a question during operant-level determination. They contributed nothing new to the conversation.

Recording. A clock, or watch with sweep-second hand, a pencil, and something to write on were necessary for the recording. One *E* was able to record the whole conversation on a tape-recorder. The *Es* ticked off each statement occurring in successive one-minute intervals by making a series of doodles incorporating marks, or by

making marks on the margin or text of a book or magazine. Different marks were used for opinions and other statements. Recording proved inconspicuous, and in only one or two cases did an *E* have to terminate an experiment because *S* seemed to notice his recording.

Although problems arose occasionally, *Es* by and large had no difficulty in arriving at and maintaining a criterion for a "sentence" or "statement," i.e., for the unit of speech that they counted, and for the subclass, statement of opinion.

The criteria varied from experimenter to experimenter, in that the rates of speaking of two subjects reported by the same *Es* are correlated, and the reported rates are a function not only of the subject's rate of speaking, and of *E*'s rate of speaking in reinforcing, but also of the criterion for "statements" adopted by *E*.

In only one case did an *S* comment on *E*'s recording: during extinction he asked *E* what she was doodling, and was satisfied when she showed him her scribbles. The *Es* also noted *S*'s general behavior during extinction, and the mode of termination of the experiment.

Execution. In a few cases, the experiment was begun, and then terminated by phone calls, third persons entering the room, or because *E* feared that *S* had noticed that he was recording. All the experiments completed are reported in this paper, except one from group A_1EA_2 , whose data could not be accurately transcribed. Under questioning, no experimenter reported that he terminated the experiment because results did not seem satisfactory to him.

Two *Es* carried out operant-level determination for only 9 minutes, and one went overtime. Four went overtime during reinforcement. The greatest variability appeared during extinction; seven *Ss* failed to continue talking for 10 minutes following the beginning of disagreement, or of nonreinforcement, either leaving the room or falling into silence. Eight *Es* carried on the conversation past the 10-minute minimum extinction period. Since *Es* were not consistent in continuing to record or to converse past this time, data are reported only on the first 10 minutes.

In summary, the experiment is designed to determine whether a person, in conversation with another person, can manipulate the second person's conversation by agreeing or disagreeing, or by paraphrasing. The experimenter himself, it should be noted, contributes nothing new to the content of the conversation.

RESULTS

Awareness. No *S* ever gave any evidence that he was "aware" that he was serving as a subject in an experiment, that his behavior was being deliberately manipulated and recorded, or that he recognized that there was anything peculiar about the conversation. The only qualification that must be made is this: during extinction, some *Ss* got angry at *E* and commented on his disagreeableness, or noted his "lack of interest," and during reconditioning one member of group A_1EA_2 gave *E* "queer, searching glances," perhaps because of the opinions that *E* was now agreeing with. These changes of behavior are consistent with those found in other situations when *S* is undergoing extinction (3).

Conditioning is demonstrated if the appropriate changes appear in the rate of speaking opinion-statements as a function of the conditions of reinforcement. When reinforcement is given, the rate must increase; when it is withdrawn, the rate must decrease.

Distributions were made of the number of opinion-statements (N_{opin}) and of all statements (N_{all}), and their cumulative values (CN_{opin} and CN_{all}) for each minute of the three experimental periods. From the latter, mean rates of making statements were computed. Relative frequencies of opinions ($RF_{\text{opin}} = CN_{\text{opin}}/CN_{\text{all}}$) were determined for each *S* for each period.

Rates. The rates of making statements (CN_{all}/t) showed no significant changes as a function of reinforcement. Table 2 gives, in

Table 2
MEDIAN AND RANGES FOR EACH 10-MINUTE PERIOD

10-Minute Period		Groups OAE, OAD, OPE, OPD Combined			Group A_1EA_2		
		Proc.	Median	Range	Proc.	Median	Range
Rate (statements/minute)	1st	op	5.3	2.2-12.8	cond	7.1	2.4-14.0
	2nd	cond	5.7	3.2-17.1	ext	6.3	1.9-11.0
	3rd	ext	5.2	1.4-12.8	recond	5.8	2.9-14.5
Relative frequency of opinion-statements	1st	op	0.320	.012-.655	cond	0.574	.208-.653
	2nd	cond	0.558	.071-.702	ext	0.302	.094-.526
	3rd	ext	0.333	.048-.643	recond	0.603	.267-.699

the upper portion, data on the distribution of these rates for each interval. Several nonparametric tests for significance of difference were made, and none showed that the null hypothesis (no difference as a function of period, manipulation, or group) could be rejected. The "priming" of *S* by means of the question, "What did you say?," seems to maintain the rates in the operant periods, and in the extinction period of group A_1EA_2 , although decreases in rate may be obscured by the fact that *E* is saying little during these times. The rank-order correlation of operant-level rates of speech obtained on two *Ss* by the same *Es* was 0.65 ($N = 14$). . . .

Relative Frequency of Opinions. Table 2 (lower portion) presents the medians and ranges of the distributions of RF_{opin} for each period. Each of the 24 *Ss* showed an increase in his relative frequency of opinion during the reinforcement period over his operant level, or (for group A_1EA_2) over his preceding extinction period. The probability that this result would have been obtained if there had been no effect of the experimental variable is $(1/2)^{24}$. Twenty-one of the 24 showed a *reduced* RF_{opin} in the extinction or disagreement period below that of the preceding period of reinforcement. The probability that fewer than four *Ss* would not change in the absence of an effect of the experimental variable is 1.1 $(1/2)^{13}$. Signed rank tests of the significance of the differences yield *p* values well below .01.

The magnitude of the effects can be evaluated by determining two ratios for each *S*: (a) that of RF_{opin} obtained during conditioning to RF_{opin} of the operant level or (for group A_1EA_2) RF_{opin} in reconditioning to RF_{opin} in the preceding extinction period, and (b) of RF_{opin} during the extinction period to RF_{opin} during the preceding conditioning period. Large values of the former of these ratios are possible only when the operant level RF_{opin} is low. Table 3 presents the mean, median, and range of these values for groupings of the 24 *Ss* based on the methods of reinforcement and extinction.

An evaluation was made of the relative effectiveness of agreement and paraphrase in conditioning, and of disagreement and silence in extinction. Fisher's exact test of independence in contingency tables was applied about the medians of Table 3A for groups OAD and OAE taken together versus OPE and OPD, and about the medians of Table 3B for groups OAD and OPD against OAE and OPE. No

Table 3

MEANS, MEDIANS, AND RANGES OF RATIO-INDEX OF CHANGES IN
RELATIVE FREQUENCY OF OPINION-STATEMENTS

Groups Combined	RF Ratios in Distribution	N	Mean	Medium	Range
<i>A. Conditioning effect (no effect: ratio-index = 1.00)</i>					
OAD, OAE	A/O	9	2.27	1.76	1.50- 5.70
A ₁ EA ₂	A ₂ /E	7	2.29	2.17	1.09- 4.32
OAD, OAE, A ₁ EA ₂	A/O, A ₂ /E	16	2.28	1.85	1.09- 5.70
OPD, OPE	P/O	8	4.23	2.02	1.05-11.47
All	A/O + A ₂ /E + P/O	24	2.91	1.85	1.05-11.47
<i>B. Extinction effect (no effect: ratio-index = 1.00)</i>					
OPE, OAE	E/P, E/A	11	0.71	0.70	0.48-0.86
A ₁ EA ₂	E/A ₁	7	0.66	0.52	0.45-1.15
OPE, OAE, A ₁ EA ₂	E/P, E/A, E/A ₁	18	0.69	0.52	0.45-1.15
OPD, OAD	D/P, D/A	6	0.65	0.62	0.27-1.01
All	E/P, E/A, E/A, D/P, D/A	24	0.67	0.65	0.27-1.15

difference in the number of cases falling above and below the medians was significant at the .05 level, although the difference between agreement and paraphrase is significant between the .05 and .10 levels.

Means and variances were also computed. An *F* test of the significance of difference in the variances of OAD and OAE and of OPD and OPE gives 8.239 ($df = 8, 7$), significant at better than the .005 level. Paraphrasing and agreement, although both effective, are not equivalent as reinforcing stimuli; paraphrasing is much more variable in its effectiveness (or perhaps the variety of statements made as paraphrases exceeded those called agreements).

The method of extinction also yielded a significant difference in variance: $F = 5.175$ ($df = 10, 5$), significant at the .05 level. Despite these differences in variance, group curves were constructed. All four groups were combined without respect to method of reinforcement or extinction. The median *N* and *CN* of opinions, and of all sentences, were then determined for each successive minute of each of the three periods. Figure 1 presents these medians for the groups OAD, OPD, OAE, and OPE, and for group A₁EA₂.

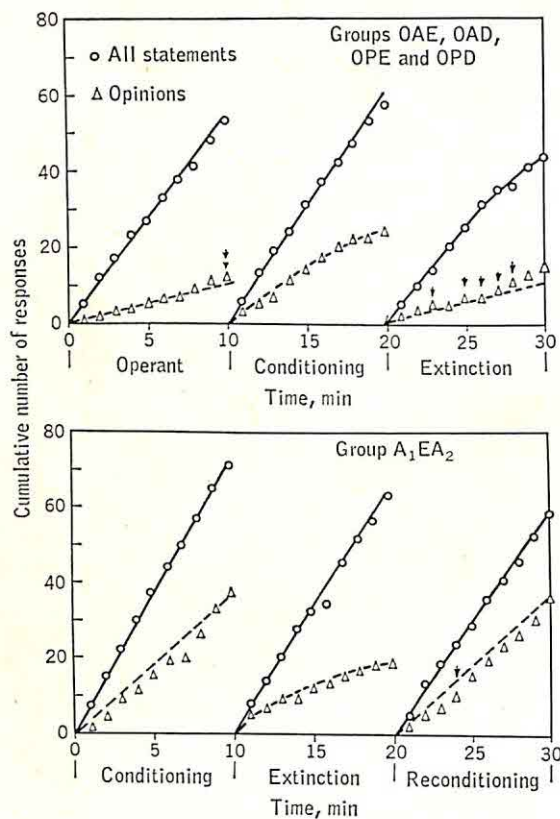


Fig. 1. Median Cumulative Frequency Curves of Opinion-Statements, and of All Statements for Each 10-minute Period of the Experiment.

For the upper graphs $N = 17$, for the lower, $N = 7$. At each arrow, N on that and successive trials is diminished by one. In the extinction period of the upper graph, each S that dropped out, "had to leave." In the other cases, E discontinued the procedure at the time indicated.

Figure 2 demonstrates that the median curves are indeed representative. In it are plotted the experimental points obtained during the operant level period from (a) the S giving a CN_{op} equaling the median, together with the S s giving (b) the lowest and (c) the highest values among the 17 S s of the combined groups, and from the corresponding S s of group A_1EA_2 , chosen about the median of the extinction period. Any other sets of individual data might have been presented, but these give some view of the spread, as well as of the consistency of results of the various subjects.

In summary, the rate of stating opinions changed in accordance with the assumptions made. All S s increased their rate of stating opinions, regardless of the topic of conversation, its setting, or S 's particular relationship with the E . The order of magnitude of the

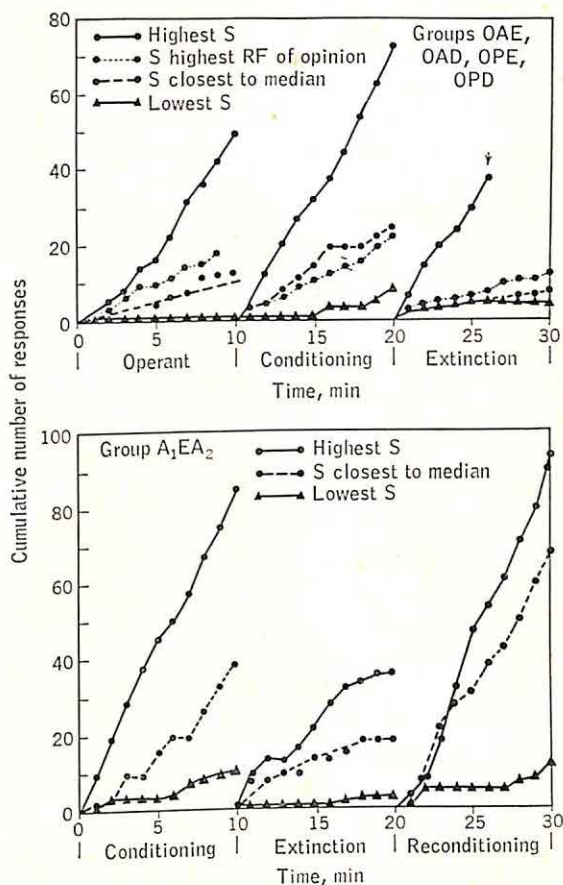


Fig. 2. Individual Cumulative Frequency Curves of Opinion-Statements for Each 10-minute Period of the Experiment, Demonstrating the Consistency of the Effect and Its Range.

effect depended upon the kind of reinforcement employed. How it may be related to the variables noted above cannot be inferred from the present data.

DISCUSSION

Individual differences in the rates of speech, and of giving opinions, are most striking and highly significant. We have already noted that they are the joint outcome of *S*'s rate of speech, the length of his sentences, of *E*'s discrimination of his speech, and of *E*'s own speech rate. Of the two *S*s with the lowest rate of making statements,

one was a Finn who spoke English with difficulty, and the other was a young woman who talked very fast and in very long sentences indeed. (She was also the most opinionated, according to our rate of giving opinions.) Since the experiment was performed, Fries's work has become available, and a study of it suggests the basis of our *Es* criteria.

The statements that the *Es* counted during the period of reinforcement are evidently identical with Fries's "utterance units," i.e., stretches of speech bounded by a change of speaker. During reinforcement and during extinction by disagreement, each stretch of *S*'s speech is bounded by *E*'s delivery of successive reinforcements or disagreements. The cues in *S*'s speech that determine *E*'s delivery of a reinforcement probably cannot yet be specified. However, the facts that the rate of uttering "statements" is stable, and that the rates reported by the same *E* are correlated with each other suggest that the "statements" or "sentences" counted during the operant level, and during extinction (although these are by definition not Fries's "utterances," since *E* says nothing) are stretches of speech such that *E* is stimulated to respond. He does so, not by speaking, but rather by making a mark in his record. If this analysis is correct, then our *S*'s statements are what Fries also terms statements, i.e., "sentences that are regularly directed to eliciting attention to continuous discourse."

Magnitude of the Effect. These data do not permit us to draw conclusions about the magnitude of the effect, although it is clearly some function of the values of reinforcement variables. If *S* rarely states an opinion, it is difficult for the number of reinforcements to become very great, and the effect is necessarily small.

Acquisition Effects. The not-quite-significant difference in the median effects of paraphrasing and of simple agreement, and the significant difference in their variances are interesting. Probably many different kinds of paraphrases were employed; the differential effectiveness of these as reinforcing stimuli needs investigation. Both the smallest and the greatest changes in the rate of stating opinions were produced by paraphrasing.

Extinction Effects. During extinction by disagreement, some *Ss* "marshalled the facts," others changed the topic. Some subjects who were extinguished by either treatment became "disturbed," or

angry. There is more than a suggestion that when *S* undergoes complete nonreinforcement, his speech tends to extinguish and, indeed, he tends to leave the experimental situation earlier ("for study," "to go to dinner," and the like), but the 10-minute extinction period is too brief, and the variation among *Es* in continuing to record is too great to permit evaluation of this tendency.

General Remarks. Certain problems, soluble by further research, set limitations on the generality of the present results.

Only one of our *Es* was able to use a tape recorder, and clearly, the use of such an instrument, perhaps in conjunction with independent judges, might yield counts of all statements and opinion-statements that were less dependent on *E*'s own criteria. However, it is not at all clear that there would be less dependence on *E*'s criteria, since the delivery of reinforcements will necessarily continue to depend on *E*'s speech habits. A variety of specific utterances by *E* were employed as reinforcing stimuli; a study of the variability in the effectiveness of various kinds of statements by *E* would be most useful.

The present results do not permit us to state how important is the particular social relationship between *S* and *E*. Would agreement by an *E* whom *S* disliked reinforce his verbal behavior? These conversations were relatively short, with the result that extinction was carried out to its asymptote in only a few *Ss*, and hence differences between the effect of disagreement and of complete nonreinforcement, although suggested, cannot be tested. Similarly, neither "satiation" effects of continuous and repeated reinforcement nor complete "talking-out" of *S* on a topic could occur. (It should be recalled that our procedure does not allow *E* to contribute anything new to the conversation.)

The topics of conversation were, in only a few cases, such that *S* might be "ego-involved" in their outcome. Perhaps if *S* were subjected to these procedures when he was talking about something he "felt deeply" about, the results might differ, e.g., acquisition might be greater and extinction far slower. Orderly changes in the topic of conversation should also be observable.

Finally, it should be remembered that our *Es* were all well trained in conditioning before undertaking this experiment, and this ex-

perience may prove necessary for the successful completion of the experiment.

Despite these limitations, this experiment shows that if, in what is ostensibly an ordinary conversation, one agrees with opinions expressed by a speaker, the speaker will give still more opinions, and that returning the speaker's words in paraphrase has the same effect. It also shows that disagreement reduces the number of opinions given, as does ignoring the speaker's statement. The verbal behavior of a speaker, apparently without regard to its content or setting, is under the control, not only of the speaker himself, but also of the person with whom he is conversing.

These results are in accord with the two hypotheses made. But one may ask, is this operant conditioning? By any empirical, non-theoretical definition of conditioning, the changes in behavior found conform with those of conditioning, and the present results may be classified as conditioning. What are some of the alternatives?

Two can be noted, and both suggest that the data depend upon the *Es'* behavior, rather than the *Ss'*. The *Es* may have "made up" the data, since they knew that certain kinds of data were expected of them. This alternative can be rejected without hesitation. The *Es'* previous performances, and the internal consistency of the data lend it no credulity. A second alternative is that "suggestion" may have altered the *Es'* discrimination of speech. If this were so, it would itself be a finding of interest. The writer is inclined to doubt very much that this occurred to any extent, in view of the phenomenon of "negative suggestibility," and of the frank skepticism of some *Es* as to the experiments' outcome before the data were collected and tabulated. Repetition of the experiment, with tape-recording of the verbal behavior of both *S* and *E* will permit ready evaluation of both these possibilities.

The results of this experiment make psychological and scientific sense of common-sense descriptions of conversation. ("People like to talk to people who are interested in what they are saying"; "if you ignore him, he'll go away"; "all right, if you don't believe me, here are the facts. . .") and, indeed, other social and political behaviors. The data suggest that, once the appropriate simplifying assumptions are made, a very high degree of order can be revealed in "complex"

situations, and that a still higher degree of order can be introduced into them.

The simplifying hypotheses made here are derived from the concepts of *response* and of *conditioning*, and they have proved experimentally fruitful in the present instance. This complex behavior is available to direct experimental investigation, and the orderliness and lawfulness of the behavior exhibits itself when irrelevant details are ignored. The heuristic advantages of much of present stimulus-response theory, when it is applied in the field of verbal behavior in a social context, are clear.

If our interpretation is correct, experimental work on a wider variety of human social behavior is possible. The isolation in conversation of independent variables susceptible to direct manipulation and of dependent variables showing orderly change, should give a much wider and more significant scope to experimental investigation. The experiments now possible provide new techniques for the investigation of client-therapist relationships and of therapeutic techniques in clinical psychology. They may be applied to the study of the behavior of small groups, and of personality.

They suggest how cooperation may be ensured. They lead to questions such as, "Can one, by pairing oneself with a reinforcing stimulus, come to control effectively the behavior of a total stranger?" That is to say, if a person agrees with everything said by someone whom he has not previously known, will he then have other means of reinforcing, or of exerting other types of control over, the stranger's behavior? The possibilities are interesting.

SUMMARY AND CONCLUSIONS

Seventeen *Es* carried on conversations with 24 different *Ss*.

Two assumptions are made, (a) that "stating an opinion" is a class of behavior that acts as a response, and (b) that statements of agreement with, or paraphrases of, such statements of a speaker act as reinforcing stimuli. From these it is inferred that the rate at which a speaker states opinions varies with the administration of agreement or of paraphrase by the person with whom he is conversing. The experimental conversations were carried out on a wide variety of topics of conversation, in a wide variety of places, and in a group of *Ss*, most of whom were college students. The expected results ap-

peared. Every S increased in his rate of speaking opinions with reinforcement by paraphrase or agreement. Twenty-one Ss decreased in rate with nonreinforcement. Over-all rates of speaking did not change significantly.

In no case was the S aware that he was the subject of an experiment, or that the conversation was an unusual one.

References

- FRIES, C. C., *The Structure of English*. New York: Harcourt, Brace and World, 1952.
- SKINNER, B. F., The generic nature of the concepts of stimulus and response. *J. gen. Psychol.*, 1935, 12, 40-65.
- VERPLANCK, W. S., The operant conditioning of human motor behavior. *Psychol. Bull.* (in press).
- WILCOXON, F., *Some Rapid Approximate Statistical Procedures*. New York: American Cyanamid Co., 1949.

DONALD J. LEWIS

Rutgers, The State University

Partial Reinforcement in a Gambling Situation*

The control of behavior through reinforcement has been discussed in the introduction to this chapter. Further research has indicated that responses which are continuously reinforced do not persist as long as responses which are intermittently or partially reinforced. This finding was tested in the investigation

* Reprinted with the permission of the author and the American Psychological Association from the article of the same title, *Journal of Experimental Psychology*, 42 (1953), 447-450.

reported here, in which partial reinforcement is used in conditioning the behavior of elementary schoolboys who are eager to win toy cowboys and toy football players. The ethics of encouraging gambling at this tender age is not considered.

The following questions may be of some interest: (1) Why did Lewis use four groups of subjects? (2) Would he have been as successful if the subjects were less motivated to win? What does this suggest about the relationship of motivation to reinforcement? (3) Why did it take longer to extinguish the response for Groups II and III? (4) How are the conditions which caused frustration here similar to those in the experiment of Waterhouse and Child? (5) Specifically, in classroom practice when might partial reinforcement be most useful? when continuous reinforcement? How would you test your hypothesis?

In the area of partial reinforcement, there have been relatively few attempts to make experimental tests of the theoretical applications of learning theory to molar human behavior. Humphreys found that the verbal expectations of human Ss reacted to partial reinforcement in much the same way as did the "expectations" of lower animals. Grosslight and Child obtained similar findings about the extinction rate of a lever-pulling response. Gilinsky and Stewart found that aspirations that are partially reinforced are more resistant to extinction than those continuously reinforced. Murphy, however, did not find the "Skinner effect" in his pinball experiment. These experiments, it should be noted, involved situations that would not be considered important to the Ss. The present experiment employed a situation that could be considered a "real" and important one to the human Ss used. In this experiment the Ss pushed buttons in a "game" with which they were presented. The response of pushing a button was rewarded continuously for some Ss, partially for others, and not at all for still others. The rate of extinction of the response was determined with reference to these conditions of learning.

METHOD

Apparatus. The apparatus was a plywood box, the front face of which supported a red and a blue light. On a small platform extending out from this face toward S were four buttons. The apparatus was wired so that the buttons, when pushed by S, made a contact that turned on either the red or blue light, but never both. Out of sight of S was a small throw switch with which E determined whether the red or blue light would go on. The rewards or reinforcements used in the experiment were toys—small plastic cowboys and football players that were purchased from a novelty company.

Experimental Design. The Ss were assigned to one of four groups randomly with Groups I, II, and III having 25 Ss each and Group IV having 20. Each group was given a different condition of reinforcement. Groups I, II, and III were given a series of 10 acquisition trials, with varying conditions of reinforcement, followed by a series of extinction trials. Group IV was never reinforced, all of its trials being extinction trials from the beginning. The design of the experiment was as follows: "x" signifies reinforcement and "o" signifies no reinforcement.

Group	Reinforcement	Acquisition										Extinction				
I	100%	1	2	3	4	5	6	7	8	9	10	1	2	3	4	...
II	50%	x	x	x	x	x	x	x	x	x	x	o	o	o	o	...
III	60%	x	x	o	o	x	o	x	o	o	x	o	o	o	o	...
IV	0%	x	x	x	o	o	o	o	x	x	x	o	o	o	o	...
		o	o	o	o	o	o	o	o	o	o	o	o	o	o	...

Subjects. All Ss were between 6.5 and 7.5 years of age. They were boys obtained from primary schools of Inglewood, California. Their IQ's were between 90 and 110. Inglewood is a fairly homogeneous middle-class community and it was assumed that most Ss were of comparable economic status. It was requested that no problem children be included among the Ss, but this could not be controlled in any other way than on the basis of the teachers' judgments. Some Ss were obtained from a summer camp in Inglewood. The same controls were exercised for these Ss as for the regular school children, and the summer camp boys were randomly mixed into the four experimental groups so that they could not bias any one group.

Procedure. The experiment was presented to S as a game that he was to play. He was given 20 toys before the first trial and was told that it was possible for him to win many more, but that he also might lose all his toys. The instructions were given orally, and were repeated until S indicated that he understood how the game was to be played. Each S was tested individually. He sat immediately in front of the apparatus, which was placed on a table with E sitting just to the side.

The instructions to *S* could not be read aloud by *E* from a card to insure standardization. This was found to be impractical because *S* neither listened nor understood. More individual attention had to be given. Still, a strong effort was made to give each *S* the same instructions, and the same attitude toward the game. The *Ss* were told:

First you take a toy out of your bag without looking in the bag. You never look in the bag or feel of it to see how many toys you have left. That would be cheating. You take the first toy you come to out of the bag and put it in this little box alongside of the game. Then you push a button, any button you care to. If the red light goes on, I reach back into this big box of toys I have, and without looking, take the first one I come to, put it in this little box and they are both yours. You take them and put them in your bag. Then you take another toy from your bag, put it in the little box and push a button. Let's pretend that this time the blue light goes on. This time I win and you lose. So I take the toy and put it in the big box and I keep it. Before you push a button you have to put a toy in the little box. You can play as long as you want. You could win a whole bagful of toys, or you could lose them all. But you can quit whenever you want. Just tell me that you want to quit and all the toys in the bag at that time are yours. Are you ready? All right, let's begin.

The *E* entered the game only to see that the game was played properly and to act as "croupier," collecting the lost toys and dispensing the ones *S* won. This also acted to some extent as a control of the time interval between trials. Although *E* did not use a stop watch, he picked up and dispensed the toys at an even rate, trying to keep the intertrial interval as uniform as possible. The interval did vary, however, for a few *Ss* took considerable time in deliberation each time before they pushed a button.

Each *S* was requested after he had finished playing not to tell anyone about the game, but undoubtedly some communication did take place. It is not felt, however, that this communication had any significant influence on the results. For one reason, no *S* ever gave evidence of discovering that the game was "rigged," or even of suspecting that it was. At intervals, some *Ss* were asked before they started to play if they had any idea about how the game worked. Most of them said "no." A few had such an unspecific idea as "you have to push a button to win a toy." Some *Ss* stated an idea about what button, or series of buttons, should be pushed in order to win, but since they all ended with a series of extinction trials, the theory was always proven to be incorrect. A few *Ss*, during the extinction series, thought that the machine was broken. When they asked, they were assured that this was not so; that the machine was functioning perfectly.

Motivation for the game was exceptionally high. Just curiosity about the game seemed to arouse considerable motivation. Then, when *S* saw the toys he could win, he became extremely eager to begin. Occasionally, an *S* would say that he did not care for the toys, or that he already had too many of them at home and did not want any more. It was found, however, that in playing the game, these *Ss* became just as excited as any of the others. Almost all *Ss* made gestures or verbal expressions of annoyance when they lost and appeared very pleased when they won.

Each *S* either won or lost according to the schedule for the experimental group into which he happened to fall. After the fifth extinction trial, *S* was reminded that he could quit when he wanted and that he could keep all the toys that were in the bag at the time of quitting. The play continued either until *S* decided to quit or until all of his toys were gone. Then *S* was asked not to tell anyone about the game. It took from 15 to 20 min. to test each *S*.

RESULTS AND DISCUSSION

Table 1 shows the relevant measures of habit strength obtained during extinction. The responses of the continuously reinforced group (Group I) were very significantly more susceptible to extinc-

Table 1
MEASURES RELEVANT TO HABIT STRENGTH AS MEASURED
BY EXTINCTION

Measure	Group			
	I	II	III	IV
No. toys at beginning of extinction	30	20	22	20
Mdn. no. responses to extinction	8.0	15.0	16.2	10.5
Mdn. no. toys <i>S</i> had at time of quitting	22	5	6	10

tion than those of either of the partially reinforced groups (Groups II and III). These differences are shown in Table 2.

The median number of responses to extinction of the *Ss* of Group IV, who never received reinforcement, was 10.5. For the continuously reinforced Group I, the median was 8.0, the difference being below that demanded for significance at the 5% level. This probably means that the response had a fairly high habit strength, for all groups, before a button was pushed. Undoubtedly previous button-

Table 2

 χ^2 DIFFERENCES BETWEEN GROUPS*

Groups	χ^2	Groups	χ^2
I & II	14.79	II & III	0.00
I & III	11.52	II & IV	3.93
I & IV	2.37	III & IV	7.17

* χ^2 of 3.84 significant at .05 level of confidence; χ^2 of 6.64 significant at .01 level of confidence.

pushing experience generalized to this situation. It should be noted that many of the partial reinforcement studies on the maze-running habit in the rat begin the partially reinforced series only after the response has acquired considerable habit strength due to continuous reinforcement. It can probably be concluded, for the present experiment, that the ten reinforcements that Group I received did not make the response perceptibly stronger, in terms of resistance to extinction, than it was before reinforcements were given. Even though the actual number of reinforcements that Groups II and III received was fewer than received by Group I, the habit strength of the partially reinforced groups was greater. The number, or amount, of reinforcements, then, would seem to be irrelevant in this situation. It seems that the pattern in which the reinforcements were administered was probably the significant variable.

Sheffield has indicated how this patterning of reinforcement achieves its effect. The superior resistance to extinction of Groups II and III of this experiment can be explained by pointing out that the button-pushing responses for both of these groups have as stimuli, before the reinforcing trials are completed, the aftereffects of both success and failure. If on one trial, for example, the button-pushing response is successful, the aftereffects of this reinforcement remain until the next trial and become part of the total stimulus for this next trial. The succeeding trial then may bring nonreinforcement. The aftereffects of this trial become part of the total stimulus for the third trial. And so on through the partially reinforced series. The continuous reinforcement series, however, is relatively homogeneous. The stimulus pattern changes very little from trial to trial. Thus, when the extinction series starts, the responses of the continuous group will be weaker due to stimulus generalization than will the responses of the partially reinforced groups. This is because the extinction stimuli are markedly different from the reinforcement stimuli for the continuous group.

For the partial groups, the stimuli do not differ as much from the reinforcement series to the extinction series.

Perhaps among the more important of the aftereffects of a non-reinforced trial is frustration. Finger has suggested that a considerable amount of frustration would be engendered by the onset of extinction. If that is true, a partially reinforced group would already have become adapted to the frustration; frustration would in fact serve as a stimulus for succeeding reinforcement or success. This would not be true for a continuously reinforced group and, possibly, the frustration for this group would lead to an emotional blocking of the response and so lead fairly quickly to extinction. Skinner has found that the extinction curve after partial reinforcement is relatively smooth, but that the extinction curve following continuous reinforcement is characterized by cyclic emotional fluctuations. Jenkins, McFann, and Clayton have found this same post-continuous cyclic fluctuation. Perhaps what the clinicians refer to when they talk about "frustration tolerance" is very similar to the effects of partial reinforcement. It is suggested that the present experimental setup would provide a situation in which frustration tolerance could be studied experimentally.

No significant difference was found between Groups II and III, the two partially reinforced groups. This was true even though Group II was reinforced 50% of the time and Group III 60% of the time. This 10% difference in the amount of reinforcement, and in the distribution of the reinforcement, had no significant effect upon the number of responses that were made during extinction.

Since not all the Ss used in this experiment had the same number of toys in their possession at the beginning of extinction, it could be conjectured that S was guided by the number of toys he had left in his bag when making his decision to quit. This was controlled in part by instructing S never to look into or feel of the bag. Furthermore, the Ss in Groups I and IV had *more* toys at the time of quitting than the Ss in Groups II and III. It seems safe to conclude that the different conditions of reinforcement were at least the sufficient conditions for the different extinction rates.

SUMMARY

The present experiment was designed to test the adequacy of the notion of partial reinforcement for a situation that could be considered a real and important one to the human Ss used. The S was presented with a "gambling" game in which he could win or lose

small toys. Four groups were used: during ten acquisition trials, one group received continuous reinforcement, one received 50% partial reinforcement, and one received 60% partial reinforcement. The fourth group was never reinforced at any time. The three groups that had received reinforcement then underwent extinction trials. It was found that the partially reinforced groups were very significantly more resistant to extinction in this situation than was the continuously reinforced group. The group that had never received reinforcement was not significantly different in resistance to extinction from the group that had been continuously reinforced during acquisition. A reinforcement theory interpretation of the results was made, and it was decided that "partial reinforcement" seemed to have a considerable degree of explanatory power on this molar level.

References

- FINGER, F. W., Retention and subsequent extinction of a simple running response following varying conditions of reinforcement. *J. exp. Psychol.*, 1942, 31, 120-134.
- GILINSKY, A. S., & STEWART, J. C., Extinction of a success operation following three conditions of reinforcement. *Amer. Psychologist*, 1949, 4, 222-223. (Abstract)
- GROSSLIGHT, J. H., & CHILD, I., Persistence as a function of previous experience of failure followed by success. *Amer. J. Psychol.*, 1947, 60, 378-387.
- HUMPHREYS, L. G., Acquisition and extinction of verbal expectations in a situation analogous to conditioning. *J. exp. Psychol.*, 1939, 25, 294-301.
- JENKINS, W. O., MCFANN, H., & CLAYTON, F. L., A methodological study of extinction following a periodic and continuous reinforcement. *J. comp. physiol. Psychol.*, 1950, 43, 155-167.
- MURPHY, M. J., The differentiation of a response: An exploratory study with human subjects. Unpublished Ph.D. dissertation, New York Univer., 1943.
- SHEFFIELD, V. F., Extinction as a function of partial reinforcement and distribution of practice. *J. exp. Psychol.*, 1949, 39, 511-525.
- SKINNER, B. F., *The Behavior of Organisms*. New York: Appleton-Century-Crofts, 1938.

B. F. SKINNER
Harvard University

Teaching Machines *

Skinner recommends teaching machines for the most humanitarian of reasons. In the following article, while paying tribute to John Dewey and the progressive educators of a by-gone era for having eliminated the harsh discipline and generally punitive practices of American education, Skinner proposes that teachers can now resume the job of educating students without returning to the birch rod. In fact, Skinner's insistence that learning can and should proceed in almost every situation and for almost every student without frustration and anxiety, and with few mistakes, is at odds with Bugelski's contention that anxiety is a necessary characteristic of all learning situations (see pp. 83-84) and with the experimental demonstration by Waterhouse and Child that frustration can have beneficial effects on performance.

Skinner's theory of operant conditioning has led to quite unique teaching-machine methods. It was this theory which prompted Skinner to prefer recall (the construction of the correct answer to a question) to recognition. The issue of recall vs. recognition is discussed at the end of this chapter by Gilbert, who shows how Skinner's approach is lacking in certain respects. Skinner's article also introduces the student to some of the jargon of researchers working on the preparation of materials for use in the machines.

May teaching machines and programmed learning promote human learning in the classroom? In answering this question, Skinner echoes the optimism voiced in his *Walden Two*. Meaningful learning will replace rote learning; the schools can return to a careful teaching of subject matter. Difficult subjects can be made easy for even the average student. Abstract materials can be learned as easily as concrete materials. Skinner's optimism about teaching most individuals highly abstract ma-

* Reprinted with permission of *Science*, 128 (October, 1958), 969-977.

terials in the sciences and mathematics contrasts sharply with the opinions and research presented in the chapters on intelligence and individual differences. Burt, for example, puts more faith in breeding than in education (pp. 434-435).

Teaching machines are not likely to replace teachers and other instructional methods in the foreseeable future. Rather they will take their place along with other audio-visual aids (see Chapter 6, pp. 000-000). The proper fitting of machines and programs into the learning activities of the classroom will become the object of much future educational research.

After reading the suggestions for programing described here by Skinner, the student should try his hand at programing material in the subject or skill area with which he is most familiar.

There are more people in the world than ever before, and a far greater part of them want an education. The demand cannot be met simply by building more schools and training more teachers. Education must become more efficient. To this end curricula must be revised and simplified, and textbooks and classroom techniques improved. In any other field a demand for increased production would have led at once to the invention of labor-saving capital equipment. Education has reached this stage very late, possibly through a misconception of its task. Thanks to the advent of television, however, the so-called audio-visual aids are being re-examined. Film projectors, television sets, phonographs, and tape recorders are finding their way into American schools and colleges.

Audio-visual aids supplement and may even supplant lectures, demonstrations, and textbooks. In doing so they serve one function of the teacher: they present material to the student and, when successful, make it so clear and interesting that the student learns. There is another function to which they contribute little or nothing. It is best seen in the productive interchange between teacher and student in the small classroom or tutorial situation. Much of that interchange has already been sacrificed in American education in order to teach large numbers of students. There is a real danger that it will be wholly obscured if use of equipment designed simply

to *present* material becomes widespread. The student is becoming more and more a mere passive receiver of instruction.

A student is "taught" in the sense that he is induced to engage in new forms of behavior and in specific forms upon specific occasions. It is not merely a matter of teaching him *what* to do; we are as much concerned with the probability that appropriate behavior will, indeed, appear at the proper time—an issue which would be classed traditionally under motivation. In education the behavior to be shaped and maintained is usually verbal, and it is to be brought under the control of both verbal and nonverbal stimuli. Fortunately, the special problems raised by verbal behavior can be submitted to a similar analysis.

If our current knowledge of the acquisition and maintenance of verbal behavior is to be applied to education, some sort of teaching machine is needed. Contingencies of reinforcement which change the behavior of lower organisms often cannot be arranged by hand; rather elaborate apparatus is needed. The human organism requires even more subtle instrumentation. An appropriate teaching machine will have several important features. The student must *compose* his response rather than select it from a set of alternatives, as in a multiple-choice self-rater. One reason for this is that we want him to recall rather than recognize—to make a response as well as see that it is right. Another reason is that effective multiple-choice material must contain plausible wrong responses, which are out of place in the delicate process of "shaping" behavior because they strengthen unwanted forms. Although it is much easier to build a machine to score multiple-choice answers than to evaluate a composed response, the technical advantage is outweighed by these and other considerations.

A second requirement of a minimal teaching machine also distinguishes it from earlier versions. In acquiring complex behavior the student must pass through a carefully designed sequence of steps, often of considerable length. Each step must be so small that it can always be taken, yet in taking it the student moves somewhat closer to fully competent behavior. The machine must make sure that these steps are taken in a carefully prescribed order.

Several machines with the required characteristics have been built and tested. Sets of separate presentations or "frames" of visual material are stored on disks, cards, or tapes. One frame is presented at

a time, adjacent frames being out of sight. In one type of machine the student composes a response by moving printed figures or letters. His setting is compared by the machine with a coded response. If the two correspond, the machine automatically presents the next frame. If they do not, the response is cleared, and another is composed. The student cannot proceed to a second step until the first has been taken. A machine of this kind is being tested in teaching spelling, arithmetic, and other subjects in the lower grades.

For more advanced students—from junior high school, say, through college—a machine which senses an arrangement of letters or figures is unnecessarily rigid in specifying form of response. Fortunately, such students may be asked to compare their responses with printed material revealed by the machine. In the machine shown in Fig. 2 [omitted], material is printed in 30 radial frames on a 12-inch disk. The student inserts the disk and closes the machine. He cannot proceed until the machine has been locked, and, once he has begun, the machine cannot be unlocked. All but a corner of one frame is visible through a window. The student writes his response on a paper strip exposed through a second opening. By lifting a lever on the front of the machine, he moves what he has written under a transparent cover and uncovers the correct response in the remaining corner of the frame. If the two responses correspond, he moves the lever horizontally. This movement punches a hole in the paper opposite his response, recording the fact that he called it correct, and alters the machine so that the frame will not appear again when the student works around the disk a second time. Whether the response was correct or not, a second frame appears when the lever is returned to its starting position. The student proceeds in this way until he has responded to all frames. He then works around the disk a second time, but only those frames appear to which he has not correctly responded. When the disk revolves without stopping, the assignment is finished. (The student is asked to repeat each frame until a correct response is made to allow for the fact that, in telling him that a response is wrong, such a machine tells him what is right.)

The machine itself, of course, does not teach. It simply brings the student into contact with the person who composed the material it presents. It is a labor-saving device because it can bring one programmer into contact with an indefinite number of students. This

may suggest mass production, but the effect upon each student is surprisingly like that of a private tutor. The comparison holds in several respects. (i) There is a constant interchange between program and student. Unlike lectures, textbooks, and the usual audio-visual aids, the machine induces sustained activity. The student is always alert and busy. (ii) Like a good tutor, the machine insists that a given point be thoroughly understood, either frame by frame or set by set, before the student moves on. Lectures, textbooks, and their mechanized equivalents, on the other hand, proceed without making sure that the student understands and easily leave him behind. (iii) Like a good tutor the machine presents just that material for which the student is ready. It asks him to take only that step which he is at the moment best equipped and most likely to take. (iv) Like a skillful tutor the machine helps the student to come up with the right answer. It does this in part through the orderly construction of the program and in part with techniques of hinting, prompting, suggesting, and so on, derived from an analysis of verbal behavior. (v) Lastly, of course, the machine, like the private tutor, reinforces the student for every correct response, using this immediate feedback not only to shape his behavior most efficiently but to maintain it in strength in a manner which the layman would describe as "holding the student's interest."

Programming Material

The success of such a machine depends on the material used in it. The task of programming a given subject is at first sight rather formidable. Many helpful techniques can be derived from a general analysis of the relevant behavioral processes, verbal and nonverbal. Specific forms of behavior are to be evoked and, through differential reinforcement, brought under the control of specific stimuli.

This is not the place for a systematic review of available techniques, or of the kind of research which may be expected to discover others. However, the machines themselves cannot be adequately described without giving a few examples of programs. We may begin with a set of frames (see Table 1) designed to teach a third- or fourth-grade pupil to spell the word *manufacture*. The six frames are presented in the order shown, and the pupil moves sliders to expose letters in the open squares.

metic does not lend itself to excerpting. The numbers 0 through 9 are generated in relation to objects, quantities, and scales. The operations of addition, subtraction, multiplication, and division are thoroughly developed before the number 10 is reached. In the course of this the pupil composes equations and expressions in a great variety of alternative forms. He completes not only $5 + 4 = \square$, but $\square + 4 = 9$, $5 \square 4 = 9$, and so on, aided in most cases by illustrative materials. No appeal is made to rote memorizing, even in the later acquisition of the tables. The student is expected to arrive at $9 \times 7 = 63$, not by memorizing it as he would memorize a line of poetry, but by putting into practice such principles as that nine times a number is the same as ten times the number minus the number (both of these being "obvious" or already well learned), that the digits in a multiple of nine add to nine, that in composing successive multiples of nine one counts backwards (*nine, eighteen, twenty-seven, thirty-six*, and so on), that nine times a single digit is a number beginning with one less than the digit (nine times *six* is *fifty something*), and possibly even that the product of two numbers separated by only one number is equal to the square of the separating number minus one (the square of eight already being familiar from a special series of frames concerned with squares).

Programs of this sort run to great length. At five or six frames per word, four grades of spelling may require 20,000 or 25,000 frames, and three or four grades of arithmetic, as many again. If these figures seem large, it is only because we are thinking of the normal contact between teacher and pupil. Admittedly, a teacher cannot supervise 10,000 or 15,000 responses made by each pupil per year. But the pupil's time is not so limited. In any case, surprisingly little time is needed. Fifteen minutes per day on a machine should suffice for each of these programs, the machines being free for other students for the rest of each day. (It is probably because traditional methods are so inefficient that we have been led to suppose that education requires such a prodigious part of a young person's day.)

A simple technique used in programming material at the high-school or college level, by means of the machine shown in Fig. 2, is exemplified in teaching a student to recite a poem. The first line is presented with several unimportant letters omitted. The student must read the line "meaningfully" and supply the missing letters.

The second, third, and fourth frames present succeeding lines in the same way. In the fifth frame the first line reappears with other letters also missing. Since the student has recently read the line, he can complete it correctly. He does the same for the second, third, and fourth lines. Subsequent frames are increasingly incomplete, and eventually—say, after 20 or 24 frames—the student reproduces all four lines without external help, and quite possibly without having made a wrong response. The technique is similar to that used in teaching spelling: responses are first controlled by a text, but this is slowly reduced (colloquially, “vanished”) until the responses can be emitted without a text, each member in a series of responses being now under the “intraverbal” control of other members.

“Vanishing” can be used in teaching other types of verbal behavior. When a student describes the geography of part of the world or the anatomy of part of the body, or names plants and animals from specimens or pictures, verbal responses are controlled by non-verbal stimuli. In setting up such behavior the student is first asked to report features of a fully labeled map, picture, or object, and the labels are then vanished. In teaching a map, for example, the machine asks the student to describe spatial relations among cities, countries, rivers, and so on, as shown on a fully labeled map. He is then asked to do the same with a map in which the names are incomplete or, possibly, lacking. Eventually he is asked to report the same relations with no map at all. If the material has been well programmed, he can do so correctly. Instruction is sometimes concerned not so much with imparting a new repertoire of verbal responses as with getting the student to describe something accurately in any available terms. The machine can “make sure the student understands” a graph, diagram, chart, or picture by asking him to identify and explain its features—correcting him, of course, whenever he is wrong.

In addition to charts, maps, graphs, models, and so on, the student may have access to auditory material. In learning to take dictation in a foreign language, for example, he selects a short passage on an indexing phonograph according to instructions given by the machine. He listens to the passage as often as necessary and then transcribes it. The machine then reveals the correct text. The student may listen to the passage again to discover the sources of any

error. The indexing phonograph may also be used with the machine to teach other language skills, as well as telegraphic code, music, speech, parts of literary and dramatic appreciation, and other subjects.

A typical program combines many of these functions. The set of frames shown in Table 2 is designed to induce the student of high-school physics to talk intelligently, and to some extent technically, about the emission of light from an incandescent source. In using the machine the student will write a word or phrase to complete a given item and then uncover the corresponding word or phrase shown here in the column at the right. The reader who wishes to get the "feel" of the material should cover the right-hand column with a card, uncovering each line only after he has completed the corresponding item.

Table 2

PART OF A PROGRAM IN HIGH-SCHOOL PHYSICS. THE MACHINE PRESENTS ONE ITEM AT A TIME. THE STUDENT COMPLETES THE ITEM AND THEN UNCOVERS THE CORRESPONDING WORD OR PHRASE SHOWN AT THE RIGHT.

Sentence to be completed	Word to be supplied
1. The important parts of a flashlight are the battery and the bulb. When we "turn on" a flashlight, we close a switch which connects the battery with the _____.	bulb
2. When we turn on a flashlight, an electric current flows through the fine wire in the _____ and causes it to grow hot.	bulb
3. When the hot wire glows brightly, we say that it gives off or sends out heat and _____.	light
4. The fine wire in the bulb is called a filament. The bulb "lights up" when the filament is heated by the passage of a (n) _____ current.	electric
5. When a weak battery produces little current, the fine wire, or _____, does not get very hot.	filament
6. A filament which is <i>less</i> hot sends out or gives off _____ light.	less
7. "Emit" means "send out." The amount of light sent out, or "emitted," by a filament depends on how _____ the filament is.	hot
8. The higher the temperature of the filament the _____ the light emitted by it.	brighter, stronger

9. If a flashlight battery is weak, the _____ in the bulb may still glow, but with only a dull red color. filament
10. The light from a very hot filament is colored yellow or white. The light from a filament which is not very hot is colored _____ red
11. A blacksmith or other metal worker sometimes makes sure that a bar of iron is heated to a "cherry red" before hammering it into shape. He uses the _____ of the light emitted by the bar to tell how hot it is. color
12. Both the color and the amount of light depend on the _____ of the emitting filament or bar. temperature
13. An object which emits light because it is hot is called "incandescent." A flashlight bulb is an incandescent source of _____ light
14. A neon tube emits light but remains cool. It is, therefore, not an incandescent _____ of light. source
15. A candle flame is hot. It is a (n) _____ source of light. incandescent
16. The hot wick of a candle gives off small pieces or particles of carbon which burn in the flame. Before or while burning, the hot particles send out, or _____, light. emit
17. A long candlewick produces a flame in which oxygen does not reach all the carbon particles. Without oxygen the particles cannot burn. Particles which do not burn rise above the flame as _____. smoke
18. We can show that there are particles of carbon in a candle flame, even when it is not smoking, by holding a piece of metal in the flame. The metal cools some of the particles before they burn, and the unburned carbon _____ collect on the metal as soot. particles
19. The particles of carbon in soot or smoke no longer emit light because they are _____ than when they were in the flame. cooler, colder
20. The reddish part of a candle flame has the same color as the filament in a flashlight with a weak battery. We might guess that the yellow or white parts of a candle flame are _____ than the reddish part. hotter
21. "Putting out" an incandescent electric light means turning off the current so that the filament grows too _____ to emit light. cold, cool
22. Setting fire to the wick of an oil lamp is called _____ the lamp. lighting
23. The sun is our principal _____ of light, as well as of heat. source
24. The sun is not only very bright but very hot. It is a powerful _____ source of light. incandescent
25. Light is a form of energy. In "emitting light" an object changes, or "converts," one form of _____ into another. energy
26. The electrical energy supplied by the battery in a flashlight is converted to _____ and _____. heat, light; light, heat

- | | |
|---|--------------|
| 27. If we leave a flashlight on, all the energy stored in the battery will finally be changed or _____ into heat and light. | converted |
| 28. The light from a candle flame comes from the _____ released by chemical changes as the candle burns. | energy |
| 29. A nearly "dead" battery may make a flashlight bulb warm to the touch, but the filament may still not be hot enough to emit light—in other words, the filament will not be _____ at that temperature. | incandescent |
| 30. Objects, such as a filament, carbon particles, or iron bars, become incandescent when heated to about 800 degrees Celsius. At that temperature they begin to _____. | emit light |
| 31. When raised to any temperature above 800 degrees Celsius, an object such as an iron bar will emit light. Although the bar may melt or vaporize, its particles will be _____ no matter how hot they get. | incandescent |
| 32. About 800 degree Celsius is the lower limit of the temperature at which particles emit light. There is no upper limit of the _____ at which emission of light occurs. | temperature |
| 33. Sunlight is _____ by very hot gases near the surface of the sun. | emitted |
| 34. Complex changes similar to an atomic explosion generate the great heat which explains the _____ of light by the sun. | emission |
| 35. Below about _____ degrees Celsius an object is not an incandescent source of light. | 800 |

Several programming techniques are exemplified by the set of frames in Table 2. Technical terms are introduced slowly. For example, the familiar term "fine wire" in frame 2 is followed by a definition of the technical term "filament" in frame 4; "filament" is then asked for in the presence of the nonscientific synonym in frame 5 and without the synonym in frame 9. In the same way "glow," "give off light," and "send out light" in early frames are followed by a definition of "emit" with a synonym in frame 7. Various inflected forms of "emit" then follow, and "emit" itself is asked for with a synonym in frame 16. It is asked for without a synonym but in a helpful phrase in frame 30, and "emitted" and "emission" are asked for without help in frames 33 and 34. The relation between temperature and amount and color of light is developed in several frames before a formal statement using the word "temperature" is asked for in frame 12. "Incandescent" is defined and used in frame 13, is used again in frame 14, and is asked for in frame 15, the student receiving a thematic prompt from the recurring phrase

"incandescent source of light." A formal prompt is supplied by "candle." In frame 25 the new response "energy" is easily evoked by the words "form of . . ." because the expression "form of energy" is used earlier in the frame. "Energy" appears again in the next two frames and is finally asked for, without aid, in frame 28. Frames 30 through 35 discuss the limiting temperatures of incandescent objects, while reviewing several kinds of sources. The figure 800 is used in three frames. Two intervening frames then permit some time to pass before the response "800" is asked for.

Unwanted responses are eliminated with special techniques. If, for example, the second sentence in frame 24 were simply "It is a(n) _____ source of light," the two "very's" would frequently lead the student to fill the blank with "strong" or a synonym thereof. This is prevented by inserting the word "powerful" to make a synonym redundant. Similarly, in frame 3 the words "heat and" pre-empt the response "heat," which would otherwise correctly fill the blank.

The net effect of such material is more than the acquisition of facts and terms. Beginning with a largely unverbilized acquaintance with flashlights, candles, and so on, the student is induced to talk about familiar events, together with a few new facts, with a fairly technical vocabulary. He applies the same terms to facts which he may never before have seen to be similar. The emission of light from an incandescent source takes shape as a topic or field of inquiry. An understanding of the subject emerges which is often quite surprising in view of the fragmentation required in item building.

It is not easy to construct such a program. Where a confusing or elliptical passage in a textbook is forgivable because it can be clarified by the teacher, machine material must be self-contained and wholly adequate. There are other reasons why textbooks, lecture outlines, and film scripts are of little help in preparing a program. They are usually not logical or developmental arrangements of material but stratagems which the authors have found successful under existing classroom conditions. The examples they give are more often chosen to hold the student's interest than to clarify terms and principles. In composing material for the machine, the programmer may go directly to the point.

A first step is to define the field. A second is to collect technical terms, facts, laws, principles, and cases. These must then be ar-

ranged in a plausible developmental order—linear if possible, branching if necessary. A mechanical arrangement, such as a card filing system, helps. The material is distributed among the frames of a program to achieve an arbitrary density. In the final composition of an item, techniques for strengthening asked-for responses and for transferring control from one variable to another are chosen from a list according to a given schedule in order to prevent the establishment of irrelevant verbal tendencies appropriate to a single technique. When one set of frames has been composed, its terms and facts are seeded mechanically among succeeding sets, where they will again be referred to in composing later items to make sure that the earlier repertoire remains active. Thus, the technical terms, facts, and examples in Table 2 have been distributed for reuse in succeeding sets on reflection, absorption, and transmission, where they are incorporated into items dealing mainly with other matters. Sets of frames for explicit review can, of course, be constructed. Further research will presumably discover other, possibly more effective, techniques. Meanwhile, it must be admitted that a considerable measure of art is needed in composing a successful program.

Whether good programming is to remain an art or to become a scientific technology, it is reassuring to know that there is a final authority—the student. An unexpected advantage of machine instruction has proved to be the feedback to the *programmer*. In the elementary school machine, provision is made for discovering which frames commonly yield wrong responses, and in the high-school and college machine the paper strips bearing written answers are available for analysis. A trial run of the first version of a program quickly reveals frames which need to be altered, or sequences which need to be lengthened. One or two revisions in the light of a few dozen responses work a great improvement. No comparable feedback is available to the lecturer, textbook writer, or maker of films. Although one text or film may seem to be better than another, it is usually impossible to say, for example, that a given sentence on a given page or a particular sequence in a film is causing trouble.

Difficult as programming is, it has its compensations. It is a salutary thing to try to guarantee a right response at every step in the presentation of a subject matter. The programmer will usually

find that he has been accustomed to leave much to the student—that he has frequently omitted essential steps and neglected to invoke relevant points. The responses made to his material may reveal surprising ambiguities. Unless he is lucky, he may find that he still has something to learn about his subject. He will almost certainly find that he needs to learn a great deal more about the behavioral changes he is trying to induce in the student. This effect of the machine in confronting the programmer with the full scope of his task may in itself produce a considerable improvement in education.

Composing a set of frames can be an exciting exercise in the analysis of knowledge. The enterprise has obvious bearings on scientific methodology. There are hopeful signs that the epistemological implications will induce experts to help in composing programs. The expert may be interested for another reason. We can scarcely ask a topflight mathematician to write a primer in second-grade arithmetic if it is to be used by the average teacher in the average classroom. But a carefully controlled machine presentation and the resulting immediacy of contact between programmer and student offer a very different prospect, which may be enough to induce those who know most about the subject to give some thought to the nature of arithmetical behavior and to the various forms in which such behavior should be set up and tested.

Can Material Be Too Easy?

The traditional teacher may view these programs with concern. He may be particularly alarmed by the effort to maximize success and minimize failure. He has found that students do not pay attention unless they are worried about the consequences of their work. The customary procedure has been to maintain the necessary anxiety by inducing errors. In recitation, the student who obviously knows the answer is not too often asked; a test item which is correctly answered by everyone is discarded as nondiscriminating; problems at the end of a section in a textbook in mathematics generally include one or two very difficult items; and so on. (The teacher-turned-programmer may be surprised to find this attitude affecting the construction of items. For example, he may find it difficult to allow an item to stand which "gives the point away." Yet if we can

solve the motivational problem with other means, what is more effective than giving a point away?) Making sure that the student knows he doesn't know is a technique concerned with motivation, not with the learning process. Machines solve the problem of motivation in other ways. There is no evidence that what is easily learned is more readily forgotten. If this should prove to be the case, retention may be guaranteed by subsequent material constructed for an equally painless review.

The standard defense of "hard" material is that we want to teach more than subject matter. The student is to be challenged and taught to "think." The argument is sometimes little more than a rationalization for a confusing presentation, but it is doubtless true that lectures and texts are often inadequate and misleading by design. But to what end? What sort of "thinking" does the student learn in struggling through difficult material? It is true that those who learn under difficult conditions are better students, but are they better because they have surmounted difficulties or do they surmount them because they are better? In the guise of teaching thinking we set difficult and confusing situations and claim credit for the students who deal with them successfully.

The trouble with deliberately making education difficult in order to teach thinking is (i) that we must remain content with the students thus selected, even though we know that they are only a small part of the potential supply of thinkers, and (ii) that we must continue to sacrifice the teaching of subject matter by renouncing effective but "easier" methods. A more sensible program is to analyze the behavior called "thinking" and produce it according to specifications. A program specifically concerned with such behavior could be composed of material already available in logic, mathematics, scientific method, and psychology. Much would doubtless be added in completing an effective program. The machine has already yielded important relevant by-products. Immediate feed-back encourages a more careful reading of programmed material than is the case in studying a text, where the consequences of attention or inattention are so long deferred that they have little effect on reading skills. The behavior involved in observing or attending to detail—as in inspecting charts and models or listening closely to recorded speech—is efficiently shaped by the contingencies arranged by the machine. And when an immediate result is in the balance, a student will be

more likely to learn how to marshal relevant material, to concentrate on specific features of a presentation, to reject irrelevant materials, to refuse the easy but wrong solution, and to tolerate indecision, all of which are involved in effective thinking.

Part of the objection to easy material is that the student will come to depend on the machine and will be less able than ever to cope with the inefficient presentations of lectures, textbooks, films, and "real life." This is indeed a problem. All good teachers must "wean" their students, and the machine is no exception. The better the teacher, the more explicit must the weaning process be. The final stages of a program must be so designed that the student no longer requires the helpful conditions arranged by the machine. This can be done in many ways—among others by using the machine to discuss material which has been studied in other forms. These are questions which can be adequately answered only by further research.

No large-scale "evaluation" of machine teaching has yet been attempted. We have so far been concerned mainly with practical problems in the design and use of machines, and with testing and revising sample programs. The machine shown in Fig. 2 [omitted] was built and tested with a grant from the Fund for the Advancement of Education. Material has been prepared and tested with the collaboration of Lloyd E. Homme, Susan R. Meyer, and James G. Holland. The self-instruction room shown in Fig. 3 [omitted] was set up under this grant. It contains ten machines and was recently used to teach part of a course in human behavior to Harvard and Radcliffe undergraduates. Nearly 200 students completed 48 disks (about 1400 frames) prepared with the collaboration of Holland. The factual core of the course was covered, corresponding to about 200 pages of the text. The median time required to finish 48 disks was $14\frac{1}{2}$ hours. The students were not examined on the material but were responsible for the text which overlapped it. Their reactions to the material and to self-instruction in general have been studied through interviews and questionnaires. Both the machines and the material are now being modified in the light of this experience, and a more explicit evaluation will then be made.

Meanwhile, it can be said that the expected advantages of machine instruction were generously confirmed. Unsuspected possibilities were revealed which are now undergoing further exploration.

Although it is less convenient to report to a self-instruction room than to pick up a textbook in one's room or elsewhere, most students felt that they had much to gain in studying by machine. Most of them worked for an hour or more with little effort, although they often felt tired afterwards, and they reported that they learned much more in less time and with less effort than in conventional ways. No attempt was made to point out the relevance of the material to crucial issues, personal or otherwise, but the students remained interested. (Indeed, one change in the reinforcing contingencies suggested by the experiment is intended to *reduce* the motivational level.) An important advantage proved to be that the student always knew where he stood, without waiting for an hour test or final examination.

Some Questions

Several questions are commonly asked when teaching machines are discussed. Cannot the results of laboratory research on learning be used in education without machines? Of course they can. They should lead to improvements in textbooks, films, and other teaching materials. Moreover, the teacher who really understands the conditions under which learning takes place will be more effective, not only in teaching subject matter but in managing the class. Nevertheless, some sort of device is necessary to arrange the subtle contingencies of reinforcement required for optimal learning if each student is to have individual attention. In nonverbal skills this is usually obvious; texts and instructor can guide the learner but they cannot arrange the final contingencies which set up skilled behavior. It is true that the verbal skills at issue here are especially dependent upon social reinforcement, but it must not be forgotten that the machine simply mediates an *essentially verbal* relation. In shaping and maintaining verbal knowledge we are not committed to the contingencies arranged through immediate personal contact.

Machines may still seem unnecessarily complex compared with other mediators such as workbooks or self-scoring test forms. Unfortunately, these alternatives are not acceptable. When material is adequately programmed, adjacent steps are often so similar that one frame reveals the response to another. Only some sort of me-

chanical presentation will make successive frames independent of each other. Moreover, in self-instruction an automatic record of the student's behavior is especially desirable, and for many purposes it should be foolproof. Simplified versions of the present machines have been found useful—for example, in the work of Ferster and Sapon, of Porter, and of Gilbert—but the mechanical and economic problems are so easily solved that a machine with greater capabilities is fully warranted.

Will machines replace teachers? On the contrary, they are capital equipment to be used by teachers to save time and labor. In assigning certain mechanizable functions to machines, the teacher emerges in his proper role as an indispensable human being. He may teach more students than heretofore—this is probably inevitable if the world-wide demand for education is to be satisfied—but he will do so in fewer hours and with fewer burdensome chores. In return for his greater productivity he can ask society to improve his economic condition.

The role of the teacher may well be changed, for machine instruction will affect several traditional practices. Students may continue to be grouped in "grades" or "classes," but it will be possible for each to proceed at his own level, advancing as rapidly as he can. The other kind of "grade" will also change its meaning. In traditional practice a *C* means that a student has a smattering of a whole course. But if machine instruction assures mastery at every stage, a grade will be useful only in showing *how far* a student has gone. *C* might mean that he is halfway through a course. Given enough time he will be able to get an *A*; and since *A* is no longer a motivating device, this is fair enough. The quick student will meanwhile have picked up *A*'s in other subjects.

Differences in ability raise other questions. A program designed for the slowest student in the school system will probably not seriously delay the fast student, who will be free to progress at his own speed. (He may profit from the full coverage by filling in unsuspected gaps in his repertoire.) If this does not prove to be the case, programs can be constructed at two or more levels, and students can be shifted from one to the other as performances dictate. If there are also differences in "types of thinking," the extra time available for machine instruction may be used to present a subject in ways

appropriate to many types. Each student will presumably retain and use those ways which he finds most useful. The kind of individual difference which arises simply because a student has missed part of an essential sequence (compare the child who has no "mathematical ability" because he was out with the measles when fractions were first taken up) will simply be eliminated.

References and Notes

- MENGER, K., "New approach to teaching intermediate mathematics," *Science* 127, 1320 (1958).
- PRESSEY, S. L., *School and Society*, 23, 586 (1926).
- , *ibid.*, 36, 934 (1932).
- SKINNER, B. F., "The experimental analysis of behavior," *Am. Scientist*, 45, 4 (1957).
- , *Verbal Behavior* (Appleton-Century-Crofts, New York, 1957).
- , "The science of learning and the art of teaching," *Harvard Educational Rev.*, 24, 2 (1954).
- , *Science and Human Behavior* (Macmillan, New York, 1953).
- Dr. Homme prepared sets of frames for teaching part of college physics (kinematics), and Mrs. Meyer has prepared and informally tested material in remedial reading and vocabulary building at the junior high school level. Others who have contributed to the development of teaching machines should be mentioned. Nathan H. Azrin cooperated with me in testing a version of a machine to teach arithmetic. C. B. Ferster and Stanley M. Sapon used a simple "machine" to teach German [see "An application of recent developments in psychology to the teaching of German," *Harvard Educational Rev.* 28, 1 (1958)]. Douglas Porter, of the Graduate School of Education at Harvard, has made an independent schoolroom test of machine instruction in spelling [see "Teaching machines," *Harvard Graduate School of Educ. Assoc. Bull.* 3, 1 (1958)]. Devra Cooper has experimented with the teaching of English composition for freshmen at the University of Kentucky. Thomas F. Gilbert, of the University of Georgia, has compared standard and machine instruction in an introductory course in psychology, and with the collaboration of J. E. Jewett has prepared material in algebra. The U.S. Naval Training Devices Center has recently contracted with the University of Pennsylvania for a study of programs relating to the machine instruction of servicemen, under the direction of Eugene H. Galanter.

The Rationale of Intrinsic Programming

The chief issue of the debate between Crowder and Skinner is the best sequencing of learning materials. Crowder takes his cue from communication theory while Skinner takes his from learning theory. Neither leaves much room for the other. Crowder does not believe that the student needs to recall the correct answer to a question and record it in the blank space provided, a practice which Skinner believes essential and consistent with the principle of reinforcement (see pp. 166-167). Crowder prefers the multiple-choice or recognition type of question originally used by Pressey. On the surface, this seems like a trivial issue. However, when one sees the different types of programs that result from the application of one or the other of these, it takes on some importance. Gilbert's paper refers to this issue also.

As we have indicated above, quite different basic assumptions have resulted in quite different practices. Crowder views teaching as a communication system which gets established between teacher and student. The student's feedback to the teacher should be the teacher's basis for deciding the future course of instruction for the student. The necessary feedback differs for different students. Crowder pays particular attention to individual differences in level and ability of learning—what Gagné and Bolles have called *readiness factors* (see pp. 36-41). Crowder believes that Skinner's "linear" programs are too rigid to accommodate individual differences. Crowder insists that learning is always remedial, in the sense of meeting in-

* Mr. Crowder is Vice President and Technical Director of the Educational Science Division of U.S. Industries, Inc., 250 Park Avenue, New York 17, New York. Published with the permission of the author.

dividual needs. In his programs, if the student makes an incorrect response, the machine very kindly and competently "washes him back" and gives him the additional help he needs.

Skinner, however, believes that individual differences are the results of gaps in learning which have occurred because certain steps have been inadvertently omitted in the student's previous learning. If we avoid such omissions and gaps, the problem which disturbs Crowder need not arise. Learning is sequential and total.

In reading this selection and considering the issues, the student might try to answer the following questions: Does Crowder place too much emphasis on individual differences while ignoring some other and perhaps more stable characteristics of learning situations? Does his approach make more complete use of other teaching methods and audio-visual aids than does Skinner's method? Does Crowder's belief that you only know how good your program is when you try it out ignore what Melton, McDonald, and others have urged about using theory as a basis for formulating practice rather than leaving practice to pure chance? Finally, is it possible that varying objectives and conditions will, at different times, show both types of programs to be useful?

The pupil-tutor relationship was the model for the auto-instructional technique known as intrinsic programming. The characteristic feature of the pupil-tutor relationship is interaction. The pupil responds to what the tutor does, and the tutor modifies his behavior on the basis of what the pupil does. The major structural features of intrinsically programmed material are designed to permit this same sort of interaction without a live tutor; the rationale of the method derives from the fact that the necessary two-way responsiveness can be achieved with practical devices.

The basic structure of intrinsically programmed material is quite simple. In each program step, the student is given a unit of material to read, usually a paragraph of thirty to seventy words. This material is followed by a multiple-choice question. The student's

answer choice determines directly and automatically what material he will see next. If he chooses the right answer to the question, he is automatically presented with the next paragraph of material and the next question. If he chooses an incorrect answer, he is automatically presented with material written specifically to correct the particular error he has just made; at the end of this correctional material the student will, in the simplest case, be directed to return to the original presentation to have a second try at the original question. However, the material at which the student arrives by making an error may be the start of a sub-program, or subsequence, of instructional material and questions in which the originally troublesome point is explained in smaller steps or with a different approach.

The student works through a subsequence just as he does through the main program, advancing when he chooses correct answers to the questions he encounters, coming to specific remedial material if he makes errors. Subsequences of any complexity desired can be prepared. When the student has worked through the subsequence, he may be returned to the point in the main sequence at which he made his initial error, to a previous step in the main sequence, or to a succeeding step in the main sequence. This is arranged at the option of the programmer and then takes place quite automatically, as a function of the answers that the student chooses.

The crucial and identifying feature of intrinsically programmed materials is the fact that the material presented each student is continuously and directly controlled by the student's performance in answering questions. To permit this step-by-step control of the program by the student, the questions are put in multiple-choice form. The choice of an answer to a multiple-choice question can be directly translated into a distinct physical act (turning to a particular page or pushing a particular button on a machine) which can then bring the appropriate material into view.

There has been a popular tendency to consider the use of the multiple-choice question as identifying intrinsically programmed material. This is an unfortunate oversimplification, since it directs the attention away from the crucial point, which is the responsiveness of the material, toward the means by which this responsiveness is achieved. A program with multiple-choice questions is not an

intrinsic program unless each separate answer choice in each question leads the student to material prepared especially for the student who has made that particular choice.

The rationale of intrinsic programming postulates that the basic learning takes place during the student's exposure to the new material on each page. The multiple-choice question is asked to find out whether the student has learned; it is not necessarily conceived as playing an active part in the primary learning process involved. The view of the learning process itself is thus essentially a naturalistic, or, if you will, naive, view. We do not pretend to know in any very useful detail exactly why students are able to learn from exposure to symbolic material, but we postulate with great confidence that learning can so occur. The same postulate underlies virtually all formal communication between human beings.

It is worthwhile to observe that this postulated ability to learn from a single exposure to symbolic material is not found in any useful degree in infra-human organisms. While it is true that a pigeon may, with some pains, be taught to discriminate between one arbitrary symbol and another, it is not true that one can say directly to a pigeon, "If you peck at this triangle we will give you a pellet, whereas if you peck at this square we will not." We can directly say much more complicated things to a human student with a high expectation that the human student will, on the first trial, perform the required task successfully. The rationale of intrinsic programming frankly relies on there being a crucial difference between the symbol-handling capacity of people and of pigeons. We attempt to make maximum use of the greater symbol-handling capacity found in people.

The direct purpose served by the questions in intrinsically programmed material is to determine whether the student has understood the material he has just read. Our reason for wanting to make this determination is that we know that the process of symbolic communication is fraught with the possibility of error, and if there has been an error, or failure of communication, we wish to detect and correct the error before proceeding. By letting the student choose the next material he will see (by his act of selecting an answer to a multiple-choice question), we have the possibility of detecting and thereby taking direct automatic action to correct any errors that occur.

We have based our technique on the possibility of detecting and correcting errors because we think it both impractical and undesirable to attempt to eliminate errors. We think it is impractical to eliminate errors because of the inevitable individual differences, both in ability and information, that will occur among our students. We think it is undesirable to eliminate errors because to do so we would have to present material in such small steps and ask such easy questions that we would not be serving the educational objectives we desire to serve.

On the question of individual differences alone it seems self-evident that a program written in such tiny steps as to allow the dullest student to succeed almost all the time must inevitably waste the time of the brighter student. It is also unrealistic to believe that in any practical situation all students come to the beginning of a program with the same amount of information. The alternatives available are (1) to make the most pessimistic possible assumptions about the background of the student and teach all of the information to all of the students or (2) to provide diagnostic questions and remedial material as an integral part of the program for those who demonstrate a need for such material. We chose the latter course.

The question of the nature of the educational objectives is more complex, but very relevant to the question of error. Much of the criticism of programmed instruction proceeds from the belief that students studying programmed material are not required or encouraged to think out a problem consciously and deliberately, or given any opportunity to stretch their mental muscles in reaching for generalizations or conclusions. It is hard to see how a technique that depends for its success on error-free performance by the student and has this objective throughout the program can avoid this criticism. The criticism is completely avoided by the intrinsic programming technique. Students studying intrinsically programmed materials can be challenged at any level we wish to set, since means are provided to assist those who do not meet the challenge successfully.

If full advantage is taken of the flexibility of intrinsically programmed material, it is possible and I think desirable to develop a program in which the error rate for the best student is not significantly lower than that for the poorest student, the better student obviously having been exposed to more difficult materials and to

many less steps. It would even be possible to structure programs so that the error rate (not the total number of errors, of course) for the bright student exceeded that encountered by the slower student, and this would, on the face of it, seem quite desirable, since the slower student may need a lower error rate for motivational purposes.

The whole question of step size and the inextricably linked question of error rate can be summed up nicely with the analogy of training hurdlers. The only way we can be sure that we are training them at a level commensurate with their present ability and which gives them some occasion to improve is if a hurdle gets knocked over now and then. As long as we are able to continuously adjust the height of the hurdles, we can achieve this optimal difficulty level. The fact that we can have a training session in which no hurdles are knocked down does not prove that we are accomplishing optimal training. It proves that the hurdles are too low.

It should be obvious that in all this discussion of error we are talking about errors that arise from making the program steps too long, assuming knowledge that the student does not yet have, or employing reasoning processes that are too subtle for the student to follow. When we realize that the alternatives—making the steps too short, reteaching knowledge the student already has, or employing reasoning less sophisticated than the student could follow—are equally undesirable, we can realize that the bald statement, "When the student makes an error, the fault is in the program" is a gross oversimplification. Legitimate errors arise from attempting to make the communication process as efficient as possible. When we have an automatically adjusting program, we can make the process efficient for each student, not merely successful for all students but grossly inefficient for most. The only legitimate errors, of course, are those that proceed from overestimating some students, and the above remarks about the desirability of some error are not meant to condone poorly written, incomplete, or unnecessarily obscure programs.

In summary, the technique of intrinsic programming postulates that the basic learning takes place by an unanalyzed process during the student's exposure to written or, in general, symbolic material. The major function served by the multiple-choice questions is to test whether the postulated learning actually took place. The test result is used to control the next material the student sees, either

advancing the student or supplying remedial material as indicated. The result is material that automatically adapts to individual differences among students and which allows us to set the difficulty level of the material and of the questions at whatever level our educational objectives and subject matter require. It is also recognized that the inclusion of the questions serves other desirable purposes, such as keeping the student active, making it clear to him what he is expected to learn from the basic material, keeping him informed of his progress, and other desirable motivational and practice purposes, but the basic purpose of the questions asked is to control the presentation of the material.

CHARLES B. FERSTER and STANLEY M. SAPON

An Application of Recent Development in Psychology to the Teaching of German^{*}

The following selection does not describe a teaching machine as such, although tape recorders are used. The sheets of paper with printed pairs of German and English words function somewhat as machine-programing. Since foreign language training has become so important, it has been necessary to take a very hard look at antiquated language curricula and to utilize fully what we are learning about language teaching and how modern technology can help.

This paper illustrates how experimental research on teaching machines and programmed learning can be carried out. The authors discuss and show how operant conditioning (see pp. 192-

^{*} Reprinted and abridged with the permission of the publisher from the article of the same title, *Harvard Educational Review*, 28 (1958), 58-69.

193) can be the basis for teaching something as complicated and as difficult as the German language. The authors have employed other learning principles as well: for example, principles of concept formation were used in making decisions about how to teach the more complicated parts of the language (see Chapter 4, pp. 209-322); the concept of level of aspiration was used to gauge the level of difficulty for each step (see Gagné and Bolles, pp. 38-39); and accepted principles of spaced practice, review, and overlearning were also carefully applied (see Gagné and Bolles, pp. 40-51).

The student should note how carefully the authors limit the interpretation of the data they obtained. This paper can serve as a model of the scientific attitude. In reading this selection the student should try to pick out those parts of the research design and the German program itself where specific learning principles, especially the principle of reinforcement, are utilized.

Description of Materials and Methods

The instructional material is composed of sheets of paper on which are printed pairs of sentences, equivalent in meaning, in German and English. The sheets are used with a mask which permits exposure of one line of material at a time as well as the scoring column, which is always visible. The student exposes a line of the material presenting an English sentence, and writes on scrap paper the German counterpart. He then exposes the next line of the instructional material, which contains the correct translation of the English sentence. If what the student has written and the printed German conform exactly, the student takes credit for a correct response and goes on to the next item. If there is any discrepancy between what the student has written and the second line of the materials, the student re-covers the second line and repeats the process. Figure 1 (see next page) is a sample of the actual instructional material which we will use to illustrate the method more concretely. Each item is composed of two lines, with English on the lower line and the German text on the *other* line. Opposite each

item are seven column spaces, which are used to record success or failure. The mask is used to expose one line of the material at a time, leaving the column to the right uncovered. When the student's answer and the text on the next line correspond exactly on the first attempt, the student places a check in the first column to the right. If there is any discrepancy, he places a zero. Additional correction attempts are not scored. The student proceeds in this manner item by item through the unit. At the end of the unit he is instructed to return to the first item in the unit and continue the process as before. Successes or failures for this second trial run are noted in the second column to the right. The student repeats this process until every English sentence has been exactly translated into German in *two consecutive* attempts. The student can scan the scoring columns at any time and determine which items he can skip because they already have two consecutive check marks, working only on those items which he has not yet mastered. The scoring columns of the sample page in Figure 1 reveal that on the first item, at the bottom of the page, the criterion was reached after four trials. On the second item, the criterion was reached in three trials, and in the third item, six trials were required. The conformance of the student's written German with the material he uncovers on the

Continue on No. 2D-1 until the criterion is met then go on to No. 3A-1						
der alte Advocat	0	0	✓	0	✓	✓
the old lawyer						
eine klein Ratte	0	✓	✓			
a small rat						
ein gutes Kind	✓	0	✓	✓		
a good child						
No. 2D-4						

Fig. 1. The Mask Is Moved Upwards, Exposing First a Line of English and Then a Line of German, and so on.

second line is the reinforcement that maintains the study behavior. The successful completion of a line of the instructional material is reinforcing because it indicates another step towards mastery of the German language.

Application of Principles in the Design of the Instructional Material

Psychological principles other than those of immediacy and the specificity of the reinforcement appear in the design of the sentences in the instructional material.

1. The amount of work per reinforcement is kept low. The occurrence of reinforcements during the course of the instructional material depends on the amount of work the student does. Reinforcements programmed on the basis of amount of work done have special effects which have been studied extensively in lower organisms. It has been found there that the organism shows a lessened disposition to return to work as the amount of behavior (work) required to obtain reinforcement is increased. This lessening of motivation continues until at extreme values the behavior will no longer be emitted. The decreased motivation is not caused by physical exhaustion, because under similar circumstances equivalent amounts of work will be done without signs of fatigue. Every time a student translates a line of the instructional material incorrectly, the amount of work he does per reinforcement is increased and his motivation is correspondingly decreased. Any factors which will reduce the number of errors, therefore, will keep the ratio of reinforcement to work high and avoid the lessening of motivation that occurs as a result of too much work per reinforcement.

2. Vocabulary. New vocabulary is introduced in a controlled manner. After each new word is introduced, it is included in the subsequent few units in order to provide practice. No attempt is made to match the natural frequency of occurrence of the various words. Words which occur very infrequently in the natural language are used as frequently as those words which occur much more frequently. This prevents overlearning of high frequency words and underlearning of low frequency words, and minimizes the amount of work required for a given amount of progress in the language. In

the construction of the material, a tally was kept of the use of each word so that all of the vocabulary was used as equally often as possible.

3. Control of overlearning. The student continues to work only on those items which he does not yet know. Items which are mastered to the required criterion are omitted.

4. Conceptual terms. All of the conceptual and more complicated parts of the language are taught without any explicit mention of the grammatical, syntactical, or conceptual principle. This is accomplished by making the student translate sentences which require the application of the principle. When the student encounters a sufficiently large number of instances in which the only *new* factor is the principle to be learned, and in which he attempts a translation in which he is variously right or wrong, he comes to behave in terms of the conceptualization even before it can be verbalized.

5. a. Graded level of difficulty. A graded level of difficulty was attempted so that the progression from item to item is so slight that the student seldom fails. In this way the student makes constant progress in the mastery of the language. Experiments from other fields in psychology show that the motivation of the learner is largely determined by the over-all frequency and the pattern of successes and failures. Ideally, as noted above, a set of materials could be constructed in which the progression from item to item is so gradual that few failures will occur.

b. Whenever a new principle, vocabulary item, or usage is being learned, it is the only thing being learned at the time. All the other parts of the material have been mastered previously. The second stage in learning these new materials is to have the student use them in varied contexts. For example, the student cannot be assumed to have a thorough mastery of the concept of the genitive case until he has worked with material that requires him to distinguish between the genitive and dative cases. This practice conforms to the basic process of concept formation in which a crucial element is an opportunity for the organism to behave inappropriately in respect to the concept. For example, if a hungry pigeon pecks at a green disc because the response is reinforced with food, we cannot be sure that he is attending to the green light until other colors are presented and the bird's pecks go unreinforced.

c. Sufficient practice and training are given so that older material is thoroughly mastered before new material is introduced. The principle of continuous mastery is particularly important in learning cumulative skills such as a foreign language. A temporary lapse of effort or attendance in an early stage of a course frequently makes it difficult, if not impossible, for a student to resume his studies with the class as a whole at a later point. This common source of difficulty is less likely to occur with the method discussed here because the rate at which the student is exposed to new material depends on his mastery of the prior material.

The Design and Administration of the Instructional Material

The instructional material contained 522 words: 286 nouns, 143 adjectives, 72 verbs, 7 adverbs and 14 prepositions. In addition to these, structural morphology was developed around the article, the interrogative, the pronoun, and the notion of the morpheme (particularly for compounding in verbal and substantive forms). In terms of syntax, the case system with its structural relations was thoroughly presented, as well as exact sentence word-order.

The content of the material closely approximated that of a first-semester college course in German. However, the net recognition vocabulary was smaller, since supplementary reading materials were not used.

The instructional material was administered to 28 subjects. . . .

The individuals who responded offered a variety of backgrounds: 4 undergraduates, 5 law students, 17 graduate students, 1 secretary, and 1 housewife. None had any prior training in German. They were informed of the purposes and nature of the investigation and were given 3 tests of aptitude for foreign-language learning. These tests have shown much promise as prognostic instruments with the adult samples in traditional courses studied by the Harvard Language Aptitude Project. The first step in instruction was a demonstration of the use of the materials, followed by a practice run in which 30 German words from the first unit were presented by the voice of a native speaker of German via tape recorder. This was done in order to forestall as much as possible gross mispronuncia-

tions, which might handicap the student's future progress in German. The subjects were then given the first part of the lesson material (an estimated 8 hours' work) and were instructed to keep an exact record of the dates and starting and finishing times for each unit of work. Subjects were free to do the work at any time or place they saw fit, and at no time were they provided with supervision or instruction by a teacher. When the initial units were completed and returned, the rest of the material was distributed.

Assessments of Results

The student mastery of German composition after completion of the instructional material was determined by his performance in the following tests:

1. A test of vocabulary on a recognition basis presenting 50 German words sampled from the material which were to be identified by a writing out of the corresponding English word. Items were scored right or wrong.
2. A test of the ability to write German sentences by the translation of English material consisting of 12 sentences totalling 100 words in German. These sentences were randomly selected from the last four (most complex) units of instructional material. Items were scored in terms of accuracy of word order, spelling, capitalization, correct article, and verb and noun endings. A failure in any one of these constituted a scorable error. For example, a word to be scored as correct must be morphologically exact, in the correct position, and spelled correctly.
3. The above also served as a measure of active vocabulary independent of structural mastery. This measurement is derived by scoring the items only in terms of the word root, ignoring syntactical considerations and minor spelling errors.

These tests, designed to measure mastery of a given corpus, were felt to have greater specific validity than currently used standardized tests, inasmuch as the latter necessarily attempt to match current textbook and classroom materials and procedures.

Inasmuch as the material represents a large and adequate sample of the German language, completion of the work sheets is in itself

a meaningful measure of accomplishment, since in reaching criteria in each unit the student has in fact written the correct German sentences from the English stimulus material.

Some estimate of the efficiency of this method of instruction is also given by the amount of time spent by the subjects. The largest individual differences occurred here.

Results

Six subjects finished the lessons in terms of achievement on content tests. The total number of hours spent on the material was reported by the subjects. The mean time spent on material was 47.5 hours. The range of scores in recognition vocabulary was 76 to 98 per cent, with a mean of 88 per cent; sentence translation 70 to 93 per cent, with a mean of 81 per cent; and active vocabulary, 90 to 100 per cent, with a mean of 96 per cent. The mean scores on the aptitude tests and the rank order of these scores compared to the ranking of subjects in terms of achievement are also shown in Table 1 (omitted). It is interesting to note that these aptitude measures which have been successfully used to predict achievement in traditional courses would seem to be much less effective in the present methodological context.

Discussion of Results

With a mean time of 47.5 hours, the six subjects learned an amount of German comparable to that presented in a first-semester course. In a semester course, however, the students spend an average of 48 hours in class with their instructor, in addition to the amount of time recommended and spent for homework. Also, the vocabulary and syntax that were learned were active in contrast to the largely passive vocabulary acquired in a conventional course.

The subjects had no instructor and were given no formal statements of grammatical principle; yet the material succeeded in teaching inductively such conceptualizations as gender, verb transitivity, morphology and syntax of the German case system, and sentence word-order.

Modifications in the Instructional Materials Dictated by the Present Results

1. LEVEL OF DIFFICULTY OF THE MATERIALS

While administering these materials, we discovered that level of difficulty was the biggest determinant of the student's sustained motivation. Whenever the level of difficulty increased too rapidly from unit to unit, so that large numbers of errors were made, the reaction was uniform in nearly all the students, who reported being emotionally upset, and who reported a strong tendency to stop even though they may have completed all of the work. What appeared before the experiment to be a very smooth progression proved to be much too difficult for the student. Those students who failed to finish the course stopped at the more difficult lessons. We assume that the uneven difficulty level of the instructional material affected the students' motivation by increasing the number of errors and hence the amount of translation the student had to do for each reinforcement. The disposition to return to the study material probably decreased as the larger number of errors produced a situation in which the student emitted too much study behavior per reinforcement to sustain further study.

2. AMOUNT OF WORK PER UNIT PROGRESS IN THE LANGUAGE

One of the problems in this particular technique of instrumenting teaching is that each line of the lessons becomes longer and contains a larger number of words as the student progresses from lesson to lesson. Therefore, the amount of gross labor involved in a given amount of progress in German increases with each lesson. This increase in the amount of work is of course not necessarily related to the new principle being taught, although it provides practice in the material already learned. In this type of teaching, however, the amount of practice is specified and designed into the material, and should not accrue haphazardly. Future modifications of the materials should include other forms, such as partial sentences in German. Here, the bulk of the sentence would already be translated;

only those parts involving the vocabulary, grammar, or usage currently being taught would be omitted.

Two points deserve further discussion here. The first, concerns the fact that the estimate of "15 to 25 hours of study" made in the posted announcements proved to be too low. Since this estimate was reached by extrapolating the results of several trial runs made with the first half of the instructional material, the effects of uneven progression in level of difficulty and accumulated overhead on working time were not apparent.

Secondly, . . . the total time spent by subject R. G. was considerably in excess of J. M., with whom she was closely ranked in terms of aptitude and achievement. Examination of the work sheets reveals that R. G.'s 78 hours were spread over a longer period of time than that taken by the other subjects, with longer intervals between work periods. There may be an implication that long lapses between work sessions have an adverse effect on economy of time without similarly affecting final achievement. This possibility suggests controlled experimentation with this type of instructional material under different conditions of spacing.

3. SHAPING NEW BEHAVIOR IN THE STUDENT

The design of the materials could be improved if each item made the correct answer in the next item more probable. The level of difficulty from item to item could be decreased by the use of all those principles of behavior which determine the form of a particular word response. For example, a series of items could be designed such that a new word never before used is made more likely to occur. The word *Fabrik* in response to the word factory could be made more probable by a preceding item such as "He saw the red fabric." Of course, later items would include the word *factory* where it was not preceded by the word *fabric*. As a *tour de force* a series of materials could probably be constructed in which each item is scientifically designed so that the student will progress from a zero knowledge of German to a complicated repertory of the level of a year of college German without ever having made an error. An achievement of this kind would be made possible through use of processes by which new verbal behavior is created rather than by the traditional processes of recall and verbal memory.

Future Course of Development

Although the present findings are only a first approximation and a preliminary instrumentation of the psychological principles being applied, the results show the possibility of teaching basic foreign language skills with high efficiency and economy of time for both student and teacher. Materials such as these could be used either for introductory pre-class instruction, so that all students could be raised to a given level of ability before they met their teacher, or they could be used as materials supplementing classroom instruction, leaving the teacher free for those tasks requiring the subtle abilities of the teacher as a human being.

THOMAS F. GILBERT
TOR Education, Inc.

On the Relevance of Laboratory Investigation of Learning to Self-Instructional Programming*

An often ignored criticism of the experimental study of human learning in the typical psychological laboratory is the fact that people can *talk*. Ignoring man's verbal capacity enables the experimenter to make some common assumptions about the behavior of rats and humans. To assume that behavior is similar throughout all species of the animal kingdom may be unwarranted. No one has proved that the assumption is valid; and as

* Reprinted and abridged with the permission of the author and publisher from the article in *Teaching Machines and Programmed Learning: A Source Book*, ed. Lumsdaine and Glaser, pp. 475-485, National Education Association, 1960.

Harlow has indicated, we may find that the central nervous system of the human being makes all the difference in the world, including the world of learning psychology.

In this article, Gilbert states that we must question the appropriateness of using both laboratory methods and concepts in the study of human learning. It is possible that some of the most crucial variables or conditions of learning are ignored in favor of those more traditionally studied in the laboratory. For example, Gilbert shows that students' behavior in answering a multiple-choice question may be much more complicated (or variable) than either Crowder or Skinner has assumed. Important variables in human behavior can easily be ignored, as he amusingly illustrates in his description of an experiment which threatened to destroy a college sophomore and an experimental psychologist, each locked in his separate world of assumptions about what was to happen.

The student should note how Gilbert's laboratory for human learning differs quite dramatically from the usual experimental laboratory. Is Gilbert suggesting that we keep colonies of culture-free human subjects as we now keep colonies of rats for experimental purposes? What does the student think about Gilbert's statement that a subject field, for example, literature, is only a way (perhaps even reprehensible way) to describe a class of *behaviors*? Do his reservations about previous laboratory study of learning contradict the views of Melton and McDonald, who see a very close relationship between the classroom and the laboratory?

I am less concerned with disclosing the irrelevance of the methodology and data of the classical human-learning laboratory than I am with prescribing a remedy. Now may be the time when prescriptions are sorely needed. This new endeavor to improve education, recently stimulated by B. F. Skinner, shows signs of growing up like Jimson weed. For evidence of uncritical and precipitate development I have to go no further than to point out that there probably exist 100 different mechanical gadgets oddly called "teaching machines"

and, in sharp contrast, maybe no more than two or three teaching programs which in any way could be called complete. Of deeper significance, I think, is the fact that there is a whole rash of so-called "control-experimental group" experiments purporting to answer questions about principles of programming education. Since I believe that there is a basis for a more considered effort, I should like to describe two alternative sets of methodological rules. The first set of rules may be useful to anyone whose foremost concern is the improvement of a given educational subject matter. The second methodology will apply to developing more general rules for programming education independently of a specific subject matter.

My qualifications for the job of providing the first set of rules are quite meager and are based on little more than a three-year-old avocation. My most reliable claim to authority in this matter is the fact that few men have seized more opportunities to make mistakes in this field than I have. These rules for getting done the immediate job of programming a specific subject matter I offer with no particular pride, but with some confidence and in dead seriousness.

RULE 1. If you don't have a gadget called a "teaching machine," don't get one. Don't buy one; don't borrow one; don't steal one. If you have such a gadget, get rid of it. Don't give it away, for someone else might use it.

This is a most practical rule, based on empirical facts from considerable observation. *If you begin with a device of any kind, you will try to develop the teaching program to fit that device.* The so-called "teaching machine" is a disease, not a challenge to self-control, and the only safe cure is to get rid of it. The recommended treatment is the cold-turkey method—don't try to taper off on programmed or scrambled textbooks.

RULE 2. Resist the temptation to design formal experiments. You don't want to know *whether* one method teaches better than another, you want to know *what* method teaches best.

This rule is based on the simple logic that a really efficient method of teaching a thing would display itself in a "control-experimental groups" study only if a really efficient method were used in the experiment. Thus you could conclude from the experiment that a method was unusually efficient only if you already knew it. In short, I am saying that the first function of this teaching laboratory is as a place of discovery, not a place to prove preconceptions.

RULE 3. Your prime purpose is to provide a student with a behavior repertory called a subject matter. If that behavior repertory is, say, physics, your problem is to take him there from whatever repertory he now has which even vaguely approximates physics.

This rule is stated to emphasize the fact that a subject matter is a class of behaviors and that everyone has some behavior which approximates that behavior class. It is easy to forget that the behaviors one goes through to master the subject matter may be different from the actual subject matter behaviors. The failure to grasp fully the implications of this rule has been, in my experience, the biggest single stumbling block for people learning to program education. The natural tendency is to begin by breaking the subject matter down into small, concise units. While this is valuable for describing the repertory you wish to build, these behavior units usually are not the ones which will actually build that repertory. They are test items, not teaching guides.

RULE 4. Get yourself an expert teacher of the subject matter you wish to program. Be wary of a college professor; he may never have seen a student learn. Remember that a good teacher is a more complicated, flexible "teaching machine" than you could possibly build. If you can't get a good program into him, you will never get one into a mechanical gadget.

This rule is not meant to suggest that the teacher is to tell you how to build the repertory. Quite to the contrary. The student will tell us this. The teacher is only the place to start.

RULE 5. Get yourself one student. I repeat, *one* student. You are about to perform an experiment in which you are permitted no degrees of freedom—that is, if the word "self" in "self-instructional" can be taken seriously. Once you have discovered an efficient program for one student, you will have described the gross anatomy of the most generally useful program.

RULE 6. You have to start somewhere, but forget that you are an expert on human learning. You aren't. Assuming you are the teacher, you should begin with the most trustworthy facilities you have available: First, trust your common sense; next, use the approximations to principles of programming that have been set down by a few people. Remember, these people probably are not more expert than you, only more audacious. They may be mostly wrong. Use their principles only as a starting place.

RULE 7. Obtain the following materials: paper, pencils, and index cards. Use no gadgets unless they are part of the subject matter. For example, if you are programming home economics, you may need an electric toaster. Resist the temptation to use the toaster as a "teaching machine."

RULE 8. You are now ready to begin programming. Think of the process as an exploratory experiment in which you do not know what the effective variables are. Your problem is to discover them. Using index cards, write out a series of questions, probes, etc., to which the student can respond. Write these items in a way that you think will lead him to a mastery of a small part of the subject matter.

It seems a good idea to write these items while having an imaginary student before you with whom you are carrying out an imaginary interchange.

RULE 9. Take your first crude effort to the student. Remember, he is going to teach you. The student cannot fail. If he doesn't get where you want him to go, you have failed. Try something else. In the absence of anything better, let whim be your guide. If you come to a dead end, vary your approach until you have gotten him where you want him to go. Tape record all sessions. The important thing to remember is to keep varying your behavior until you are successful and to describe what you do.

RULE 10. Once you have learned how to get the student through part of the material, keep going. Pay the teacher and student so they won't leave you. This can be dull work. Don't invest much time in constructing materials before you have tried them out on the student.

RULE 11. Once the teacher really appreciates immediate reinforcement and has discovered that the student alone can tell him how to teach and once he has learned how to keep varying his approach, he is more of an expert than you. Remember, it is easier to teach a physics teacher what you know about programming than it is for him to teach you physics.

RULE 12. Take your time. Education has been waiting for you since the dawn of history. When the student has the repertory you wanted to build, and when you can describe how he got that repertory, you are ready for the next step. Edit the material and try it on another student. Make whatever changes necessary for your pro-

gram to take care of both students. After fewer than 10 tries, you will have a program which will teach 98 per cent of the students. And you will have discovered how to adjust the program for individual differences.

RULE 13. Don't be too concerned that your program is not perfect. It works. It can always be revised. If you have followed the rules . . . you have done a respectable job. You can describe an exact procedure for getting most students from oblivion to mastery—not to a C average. Now, and only now, you are ready to think about automation.

RULE 14. Prepare to automate the program by discarding any "teaching machines" which, in your weakness, you kept around. Remember that you have a teacher who is a vastly complex machine and you have discovered how to make him work with efficiency. All you need to do now is to substitute more economical devices for the teacher's operations wherever you can. You probably will end with several devices. Examine each operation and fit a device to it. Never let the device dictate the program.

Suppose you have programmed physics. In examining how the teacher builds a masterful repertory in these students, you may see that he came to teach the mechanics by verbal discourse which you could simulate on a device similar to Skinner's "teaching machine" or on a paper moved under a cardboard mask. Perhaps the only possible substitute for the teacher's necessary demonstrations of certain electrical phenomena is a motion picture film or television. The compromise between efficiency and economy is your guide. There is no reason to build a machine to do what a cardboard mask or programmed-text would do as well. If it becomes evident that there are things which can be done only by a human, without resorting to excessive expense, then you simply have to write an exacting program for that human. If the program calls for an inspiring lecture, and the expense is justified, include such an inspiring lecturer amongst your instruments. After all, we are programming the subject matter, not fitting the subject matter to some preconceived machine or method. It is an idle question whether "teaching machines" or television can teach. The real and answerable question is, "How can we teach?" and the rest is a matter of classroom economics.

These rules are designed to guide one who wishes to apply whatever we already know to the task of engineering a specific educational problem. If we wish to discover principles which are not bound to a particular problem, and are still useful to the problems of education, we must resort to a much more expensive and elegant laboratory. There is no middle ground.

This more elegant laboratory we will approach with a very similar logic if we wish to provide a generalizable and relevant set of human-learning principles. The logic I refer to is that of operant conditioning methodology, and it has been described elsewhere, notably by B. F. Skinner. It deserves restatement in the context of the study of human learning.

The first characteristic of the human operant conditioning laboratory is the provision for intensive, long-term study of a single human in an artificial and fully controlled environment. This logic is based on the universal agreement that the human comes to the laboratory with a complex conditioning history. The traditional approach has attempted to get behind idiosyncrasy by statistical averages of short behavior samples. What this gives us instead of generalizable and timeless principles is an average of cultural effects. The traditional laboratory is too artificial for an interesting sociology, too sociologically sensitive for a culture-free (diachronic) behaviorism. Students may prove to use the recognition "method" better on the average simply because they have been trained that way, on the average. The diachronic laboratory, I am suggesting, will have subjects spend a great part of each day for many months in an artificial environment, hoping that environment and its associated consequences will eventually take on a unique character which will override many of the effects of conditioning in other situations. In short, we wish to make this laboratory a unique and standard culture.

Secondly, we use no, or minimal, instructions. Instructions are used as a means of getting the experiment under better control. They serve as extremely complex reinforcement conditions—the subject follows them because he fears the instructor, or wants to impress him, or feels kindly toward him; who knows? They are compounded out of an intricate sociological past. Instructions bias the experiment by telling the subject what to do. We conceive of that subject as a part of nature who is never "wrong"—and who is to instruct us

in his ways. Instructions would be standard only if they had describable and similar effects on every subject.

For purposes of control we use precisely manipulable consequences, the reinforcement value of which can be demonstrated . . . nickels, food, etc. Our subjects earn about three dollars an hour; and if they don't quit when the money is removed, they may be working to make fools of us. Instead of instructing a subject to behave in a certain way, we make his nickels contingent upon that behavior. The experimenter stays out of the experimental room. We don't know nearly enough to describe the effects of his presence.

We make each subject his own control, and we can do this by virtue of the long-term investigations. We collect months of baseline data while standardizing and making unique the experimental space. A single subject may work in this laboratory for five or more years.

While the logic of this laboratory requires it to be an artificial environment, the environment is not arbitrarily constructed. Every event in it has its own logic, the primary requirement being formal representativeness. The tasks of the traditional laboratory are chosen for traditional reasons, the nonsense syllables because Ebbinghaus used them. There are many tasks which represent serial behavior functions far freer from cultural associations than these syllables. In the diachronic laboratory we do not wish to represent specific cultural forms. A truly diachronic laboratory will allow us to obtain data from Fiji Islanders which would have the same meaning as data obtained from New Yorkers. Artificial tasks with near-zero past associations are impractical unless you can work with your subjects for many hours. The education of a single human takes many years. The scientific knowledge relevant to this education will not be gained by one-hour sessions with subjects in a laboratory hardly more rigorously contrived than the classroom itself.

The most important turnabout in the traditional laboratory logic is the conception of the very purpose of that laboratory. It is usually conceived as a place where we put to test the effects of variables chosen on a priori grounds. The a priori choice is sometimes dictated by a theory; more often, by traditional and economical considerations. What we have called "experimental design" is more accurately designated as "pre-experimental design." This a priori

selection of variables so biases our conception of nature that we sometimes throw out a subject who doesn't behave in accordance with preconception or statistical requirement. The rare subject who learns the paired-associates in one trial is the fellow who gets tossed out, either because he threw no light on some mediational hypothesis or because he violated the normal distribution. I would suggest that we would learn more if we threw out all subjects except this one.

This conception of the laboratory as first of all a testing ground biases the experimenter to ask such questions as: "Is this variable more effective than that one?" or, "Can television teach?" As a result, the laboratory is not seen as a place for discovery. It is the discovery role which will lead us to emphasize such questions as, "What variable is effective?" and "What can teach?"—these being the questions relevant to education. Under such logic, "experimental design" becomes the design of a laboratory which will allow us to get all possible "recognition" behaviors overtly expressed.

I can best demonstrate what we stand to lose when we de-emphasize the discovery function of the laboratory by relating to you an experience of a friend. This friend tells the story of his last venture into the laboratory investigation of human learning when some years ago, as a newly produced experimental psychologist, he set about to study human avoidance conditioning, presumably because there was an old hand-shock grid cluttering up the laboratory. He was well equipped in the classical tradition. He had a theory (something about an inscrutable mediating between light and shock). He had a group of male college sophomores, carefully matched for sex and education. He had an experimental design, a properly conservative two-tailed t test. And to complete the Galilean picture, he had a standardized source of uncontrolled variation: a wholly mysterious set of instructions designed to startle the most self-satisfied mediator out of its hypothetical existence. The first of the 60 scheduled sophomore proved too unsophisticated for such a high-level scientific setting. When the experimenter threw the switch which flashed a light and pulsed a shock, the subject blandly kept his hand on the grid. Trial after trial the shock intensity was increased without a twitch from the indifferent hand when, on the final trial, with the current at a searing maximum, the subject's

steady fingers began to smoke. At this point, the experimenter's fine consideration for scientific control gave way to understandable human curiosity, and he asked the subject why he didn't move his hand. The subject's reply provided perhaps the finest opportunity for scientific discovery since Skinner's first automatic food-magazine broke down. The subject simply said, "I thought you were studying how much *guts* I had!" My friend thereupon cancelled the schedules of the remaining 59 sophomores, and today he is productively engaged in the investigation of the eating and defecating behavior of hooded rats. Alas, he may have foregone the opportunity to provide us the first systematic and quantitative methodology for the study of human courage.

I relay this story in all seriousness because it served to illustrate my thesis. In this not unreasonably overdrawn example, the investigator may have failed to reduce guts to a science, but he thankfully learned in a hurry a fundamental which it is taking, or has taken, many of us much longer to learn. That fundamental is that we are not going to increase significantly our already considerable market-place knowledge of human behavior with the pedestrian methodology prevalent today. This methodology was unsystematically pieced together by the sheer inertia of tradition, the peculiar economics of college professing, and an indecisive eclecticism.

Human Problem Solving: To Find Out Is To Learn

Introduction

Most of the basic problems of psychological research are also the traditional concern of Western philosophy. Yet scientific psychology seldom acknowledges its origin in a philosophical tradition that is based on introspection more than on public evidence. Scientific psychology has, in fact, led the rebellion against the traditional view of man and his behavior. Psychological investigations give answers that are always statements of probabilities, not dogma. Particularly in the behavioral sciences, there have been periods of impatience with both the methods and the tentativeness of the conclusions arrived at by using them; the intellectual value of intuition and introspection in arriving at the "full truth" is then reasserted. This tendency to return to the methods of the earlier, philosophical psychology is especially pronounced in the area of thinking problem solving. In fact, in the next chapter, an educational philosopher (Smith) decries the way thinking has been "psychologized" (pp. 356-366).

Research in the area of problem solving is fraught with many difficulties. Duncan has pointed out most of the important ones (pp. 212-254). It is interesting, however, to note that the trend of current research is to consider human problem solving as a form of conditioning. Most investigators no longer observe the distinction between "higher" and "lower" mental processes. Illustrative of this trend is the study of creativity by Maltzman and his associates in which an operant conditioning model is used (pp. 257-310).

Duncan has suggested that research on problem-solving processes is premature (p. 241) and that the more profitable investigations may be those which deal with the relationship of learning situations and tasks to products and performances (pp. 216-235). In the discussion of programed learning, we referred to the new interest in analyzing educational tasks in order to identify their chief characteristics (p. 137). The classification of education tasks is discussed in the introduction to the article on Bloom's *Taxonomy* (pp. 569-575). It appears that both research and educational practice can be improved by such analysis and classification. Much of the research of Bruner and his associates is labeled a study of "process." These investigations however, have also been concerned with the "structure" of the learning materials, the nature and amount of guidance given to the learner, and the sequential arrangement of concepts in the curriculum.

At the present time the schools can scarcely argue that teaching bodies of loosely organized information to passively accepting students fulfills their important intellectual role. We have come to accept transition and change as a permanent cultural characteristic. We therefore no longer expect to be able to predict the future problems our students will face and the information they may find useful. The teacher who restricts his teaching to the imparting of "basic information" may be deluding himself; he may convey to students the attitude that there are more right answers than perplexing questions. Although this is not to suggest that schools must abdicate their intellectual responsibilities, it may mean, as Bruner has pointed out, that teachers should look for what is really basic in the curriculum.

To those who have been long on the education scene, these references to problem solving will not seem new. For them the image of John Dewey, the great philosopher of American education, will be quickly recalled. Among other things, Dewey lamented the school's traditional disdain for the "practical" responsibilities of its students and the passivity of students whose learning was confined to monotonous memory drills and to the recitation of definitions they would never use and laws of the physical universe they would never understand. Dewey's protests, though frequent and widely disseminated, have yet to be heard in many classrooms and

college lecture halls where the student's quest for knowledge is confined to the mindless movement of a ball-point pen across a sheet of paper. Such practice in our schools may have developed several generations of Bruner's "potshotters"—students whose responses to questions or problems are academic stabs-in-the-dark. They are forced to resort to primitive trial-and-error because they have not developed (so the hypothesis goes) more systematic and productive ways of thinking. Dewey's description of problem-solving methods was his philosophical attempt to remedy the deficiencies of rote memorization. The readings in this chapter suggest that the modern psychologist has come to Dewey's support by attempting empirically to specify those conditions which are most conducive to discovery and increased learning.

When, as teachers, we try to find out how to present the concepts of "nation" or "democracy" so that students will recognize, at some distant and yet unforeseeable moment where they can be applied, as well as their correct and incorrect applications, we are concerned with the question of *transfer of training* (or learning). This question is important in most of the articles which follow. In ordinary laboratory research the matter has been largely ignored because the experimental psychologist is more concerned with present than future performance (see Gagné and Bolles, pp. 31-32). Little educational research has been done on the problem of transfer, largely because of the large number of factors which influence it, as well as the time and expense involved in doing follow-up studies on students beyond the high school years. Gagné and Bolles discuss the "associative and intratrial factors" which may determine the successful transfer of training (pp. 41-47).

The Relationship of Readings in Chapter 4

The article by Duncan is a critical review of the research on human problem solving. It serves as a frame of reference for the succeeding readings in this chapter, especially by pointing out some of their methodological limitations. The article by Bruner concerns the use of "discovery" in teaching and suggests that greater understanding, retention, and general transfer will result with its use. The Kendler study shows how very young children may make dis-

coveries or inferences. Kersh's article is a research report on the use of a discovery method of training and indicates that there may be failures and disadvantages which Bruner has overlooked. The study by Maltzman and his associates shows how originality can be studied in terms of concrete conditions and behavior. Finally, the investigation of Kopstein casts considerable doubt on the value of repeated trials and distributed practice—traditional variables that are considered in much research in learning and in almost every text in educational psychology.

CARL P. DUNCAN
Northwestern University

Recent Research on Human Problem Solving*

Rote memorization, never very popular with students, has become increasingly unpopular with teachers. It is a classroom practice which even the traditionalists more often defend than adopt. Dewey, who eloquently expressed the teachers' discontent with rote memorization, strongly supported teaching methods which encourage "understanding" and "discovery." He believed that what students will ultimately find useful is not a collection of unrelated facts but skill and ability in problem solving.

Dewey's approach to learning, however, raises many questions still in need of answers: What are the previous conditions of teaching and practice which influence how well we later solve problems? What are the best concurrent conditions for problem solving? How do the individual characteristics of the problem solver influence his behavior in solving a problem?

* Reprinted and abridged with the permission of the author and the American Psychological Association from the article of the same title, *Psychological Bulletin*, 56 (1956), 397-429.

Duncan's review of studies in this area does not purport to answer these questions fully, but it does faithfully summarize the important investigations of recent years.

In reading this review, the student should consider what his answers would be to the following questions: (1) Does Duncan believe that the question of rote versus discovery methods has been answered empirically? What are his conclusions about this issue? (2) Wittrock's study (pp. 107-117) investigated the influence of set on student teaching and pupil achievement; it was suggested that set was used partly as a motivational device. Duncan discusses here the influence of set on problem solving. What do you suppose a general learning set (learning how to learn) would consist of? How could this general set help students avoid functional fixedness, as explained here? As a teacher, what, specifically, would you do to help students develop a general set for learning? (3) Duncan summarizes the available evidence on the effect of concurrent conditions on the behavior of problem solvers. How concrete should you make the problems for students you plan to teach? How difficult will you make them? What hints and aids in the area you plan to teach would you supply students? (4) The reviewer has entertained various hopes for future research. How could some of this research be carried on in the schools? Can you think of any particular research proposals? (5) Why does Duncan discourage research on problem-solving processes? Does this conflict with Bruner's views?

The present review summarizes most studies of human problem solving that were published in the period 1946 through 1957. A complete review of the literature on human problem solving would have to include studies in which problem solving tasks were used in the subject variable of rigidity. Since Chown (1959) has recently published an extensive review of rigidity, the studies of problem solving which are cited in her paper will not be covered here. In the case of topics where Chown has summarized some of the relevant studies, her paper will be cited along with other pertinent investigations.

Within the area of thinking, the present review covers only experimental and theoretical studies that dealt with the problem solving performances of normal human adults. Thus, the scope of the paper is narrower than that of other recent reviews (Humphrey, 1951; Johnson: 1950, 1955; Russell, 1956; van de Geer, 1957; Vinacke, 1952).

Definitions

Attempts to define thinking in general or problem solving in particular appear most clearly in the writings of Humphrey (1951), Johnson (1955), Maltzman (1955), Ray (1955), Russell (1956), Underwood (1952), van de Geer (1957), and Vinacke (1952). The defining characteristics most frequently mentioned are the integration and organization of past experience when the definition refers to all of thinking, and the dimension of discovery of correct response when reference is made to problem solving specifically. Problem solving is considered to be fairly high on the discovery dimension, as one way of distinguishing it from conditioning and rote learning which are presumed to involve relatively little response discovery. Underwood (1952) gives three methods for determining the amount of overlap between conditioning and thinking.

It is of interest to note that nearly all writers concerned with definitions emphasized that they were trying to define thinking or problem solving in such a way as to relate them to, not separate them from, simpler processes like learning or perception. Maltzman (1955) and a few others distinguish between productive and reproductive processes within thinking, but apparently no one any longer seriously defends a sharp distinction between higher and lower mental processes, particularly between thinking and learning. That issues of this kind are not completely dead, however, is indicated by the fact that van de Geer (1957) attempted to destroy the productive-reproductive distinction, and several other writers who were not primarily concerned with definitional problems also felt it necessary to state that thinking is part of learning or association (Cofer, 1957; Judson & Cofer, 1956; Judson, Cofer, & Gelfand, 1956; Saugstad, 1957; Weaver & Madden, 1949).

A few other writers who have been somewhat concerned with problems of definition may be mentioned. In an extensive study of

categorization and concept formation, Bruner, Goodnow, and Austin (1956) described broad classes of equivalence categories, one of which was "functional" categorization. The authors think this category includes at least those problem solving tasks where *S* must categorize an object as fitting a certain function, e.g., the pliers as a pendulum weight in the two-string problem. They also suggested that defining attributes are sometimes combined to create either new categories or empty categories, and that these types of combination often occur problem solving. These brief suggestions are worth noting if only because they represent one of the few attempts in the literature to relate two major areas of thinking research, problem solving and concept formation, other than by means of an all-inclusive definition (see also Underwood, 1952).

Galanter and Gerstenhaber (1956) define thinking in a way that seems to differ sharply from the usual definitions (although the reviewer does not really understand their position), and Maltzman's (1955) definition restricts thinking to articulate organisms. However, disagreement on definitions of either thinking or problem solving is less than might be expected; at least it was possible to hold a conference on human problem solving where some areas of agreement were evident in the absence of a definition of the field (Hovland & Kendler, 1955).

Any further pursuing of the issue of a definition of problem solving would lead into discussion of processes in problem solving behavior, or into theory. Both of these topics can be handled better after the bulk of the empirical studies has been presented.

Most of the remainder of the paper is a review of empirical studies of human problem solving. Insofar as possible, the review is organized in terms of the independent variables that influence problem solving performance. The categorization of these variables that was finally decided on is not very satisfactory. In many cases investigators used highly specific variables or conditions and failed to suggest any similarity between their variables and those used by other investigators. Thus, the reviewer's categories are necessarily too arbitrary.

Most of the studies to be reviewed seemed to fall into one of three major classes. In the first, the independent variables were introduced prior to testing on the final problem solving task, which task was the same for, and was presented under constant conditions to all *Ss*.

These studies used what is essentially a training and transfer design. In the second group, the independent variables were introduced during work on the test problem, or were changes in the problem itself. The third group contains studies where the variables were certain characteristics of the Ss used. Some attempt will be made to differentiate subclasses of variables within each of these major groups. Other papers to be covered include studies of individual vs. group problem solving, research on problem solving processes, and contributions to theory.

Transfer Following Variations in Training

DIFFERENT METHODS OF TRAINING

Methods of "Understanding." The first four studies reviewed here are similar in that all dealt more or less with transfer following training by memorization vs. training by various "understanding" methods. Hilgard, Irvine, and Whipple (1953), Hilgard, Edgren, and Irvine (1954), and Crannell (1956), all used Katona card tricks (Katona, 1940) as tasks; Forgas and Schwartz (1957) used various arrangements of letters. In all studies, different groups of Ss were first trained on problems solvable either by memorization (of, e.g., a certain order of cards or letters), or by learning, via one or more understanding methods, a principle or technique presumably applicable to many such tasks. The differently-trained groups were then tested for recall of training problems, and for transfer to both simple and difficult new tasks.

Hilgard et al. (both studies), and Crannell reported little or no difference among methods on recall or on simple transfer tasks, but on more difficult transfer tasks certain understanding methods, particularly the "Katona diagram," produced superior performance. Forgas and Schwartz found that both demonstration and discovery of the principle led to better performance than did memorization on all three tests.

All four of the studies tested the same Ss successively on all three types of tests, so the results may have been affected by differential transfer effects among tests, or by varying interactions between particular training methods and particular tests. Also, the various train-

ing methods were complex, unanalyzed variables that are difficult to evaluate. For example, a principle may be a single item, such as a formula, in which case it is easily learned by an understanding group; Forgas and Schwartz found that their memorization group, which had to learn a series of items, required about twice as much practice in original training as did understanding groups. In contrast, the Katona diagram used by Hilgard et al. and Crannell, was an understanding method that required much original training. Further, an understanding method may yield either positive or negative transfer depending on the particular test task; something like this apparently occurred with the "working backwards" method used by Hilgard et al. (1954), and Crannell.

Hilgard et al. pointed to limited understanding of even an understanding method as a source of error. The same point was made by Burack and Moos (1956), who found little transfer to solution of a mechanical puzzle from either verbal or actual presentation of illustrations of centrifugal force, and by Székely (1950a) with problems requiring use of hydrostatic principles.

The point raised above that results may depend on the interaction between a particular training method and a particular test task is illustrated in Corman's (1957) study. Groups given varying amounts of information on how to attack Katona match problems produced more solutions than groups given varying amounts of information about the principle underlying all problems, whereas the latter groups did best when tested for ability to verbalize the principle. The problem of interpreting results when Ss are tested successively on a series of tasks was also noted. Although the training variables appeared to have significant effects on training, and on simple and complex transfer tasks, effects on both types of transfer tasks disappeared when number of training problems attempted and solved were partialled out.

Székely (1950b) reported that Ss trained on the principle of moment of inertia by a "modern" method (first predict and watch demonstration of movements of a torsion pendulum, then read textbook material on mechanics) did better on the two-spheres problem, which requires application of the principle, than did "traditional" method Ss (read text, then watch demonstration). But Maltzman, Eisman, and Brooks (1956) failed to duplicate this finding. Either

method, or a combination of the methods, produced more solutions than a control group with no training, but there were no significant differences among the three experimental groups.

Craig (1956) had Ss cross out the word that was unrelated to four other words; each such group of words utilized a different principle. The Ss who were told the principle applying to each block learned more during training than uninformed Ss, and, probably because of differential learning, retained more after 31 days. But on transfer to new items the groups did not differ, although both did better than they had on training items.

In Buswell's (1956) study of patterns of thinking, one experiment concerned the discovery of generalizations. The Ss were to discover a rule whereby they could get sums of ordered columns of numbers without simply adding. The Ss found the problem difficult, and had trouble verbalizing the rule. On a test for transfer to similar problems, about half the Ss showed transfer.

Other Methods of Training. Ray (1957) required Ss to state verbally what they were going to do before they were allowed to make motor responses to a problem requiring turning off a light with switches. This verbal work facilitated problem solution, probably because, as was also shown, the verbal work increased S's tendency to respond systematically to elements of the problem.

A specific systematic approach, the half-split technique, was taught to Ss by Goldbeck, Bernstein, Hillix, and Marx (1957); the technique was to be applied in a complex lights-and-switches apparatus problem. The technique was not particularly effective until Ss were first taught the deductive skill of locating the elements of the problem to which a systematic approach could be applied; then the technique, as a device to improve efficiency, was an aid.

Kendler and Kendler (1956) reported that 3 to 4-yr.-old children could make a correct inference when their training had included all of the separate part-tasks needed to make the inference. It is possible, however, that the children used body orientation cues, in part, to make the correct response. The Ss had to change their position with respect to the apparatus in order to learn one (presumably the crucial one) of the necessary part-tasks. Significant inferential behavior occurred only when this part-task was the last one learned, i.e., immediately before the test trial (see pp. 271-277).

The preceding studies of various methods of training did not yield particularly clear-cut results. However, the studies varied greatly from one another, and most dealt with relatively unanalyzed situations. Since anything one learns probably has both positive and negative transferring effects, depending on the situation, more information is needed about what specific responses are learned under a particular training method, and what responses are required on a particular transfer task. Also, more attention should be paid to the amount and breadth of training, since a particular training method may yield positive transfer only if it is well learned in all its aspects.

AMOUNT OF TRAINING

Except in studies of set (see later), very little research has been directly concerned with the effect of degree of original learning of responses which are expected to influence problem performance. This is surprising, since in most other learning situations both positive and negative transfer effects on a task are considerably influenced by variations in amount of practice on a similar training task.

Three weeks prior to the problem solving session Marks (1951) gave some of his *Ss* a lecture which emphasized analysis of a problem into its elements. The lectured group was not clearly better than the nonlectured group on a problem requiring finding errors in square roots, although a finer method of scoring solutions produced data indicating some superiority of the lectured group.

French (1954) gave one group some preliminary training on a simpler version of a problem requiring turning off lights with buttons. This training greatly improved performance on the final problem. More importantly, training interacted significantly with length-difficulty of the problem. With no prior training, the simple 4-item problem was much easier to solve and learn than the 6-, 8-, or 10-item problems, which were clustered. But after training, the 4-, 6-, and 8-item problems were all solved about equally well, while 10 items were still quite difficult. This shift in relative difficulty as a function of prior training is an important finding that should be followed up. It also illustrates once again the point that results may depend upon interactions between particular training methods and particular transfer tasks.

Fattu and Mech (1953a) did one of the two experiments in which more than two amounts of training were employed. They compared groups given none, some, or much information about locating malfunctions in a gear train. Performance increased directly as a function of amount of training. Sato (1953) also compared groups given none, some, or much prior training with the characteristics of visual stimuli which were arranged in certain ways to provide problems. Difficulty of the problems was also varied. In general, differences in amount of training were significant for child Ss, but problem difficulty was more important for adult Ss.

Although they performed no experiments, Bloom and Broder's (1950) work suggests that problem solving proficiency may be improved by general training that is not tied to particular kinds of problems. A general approach to problems (essentially a checklist) was developed from comparisons of the problem solving behavior of high grading and failing college students. Training sessions with the checklist improved performance of failing students on various examinations, although control groups were not employed. Bloom and Broder's laboriously developed checklist deserves further study; there were hints in their work that training with the checklist might transfer positively to a wide variety of problems.

None of these studies of amount of training varied some reasonably unidimensional method of training over a wide range. Even in the Fattu and Mech and Sato experiments, "much" training involved a qualitative as well as a quantitative change over "some" training. Only in research on set can one find a study relating degree of original learning, systematically varied, to amount of transfer.

SET

Some situations are problems for adult Ss not because of deficiencies in S's intelligence, motivation, or past experience, but because S is set to respond in certain ways. These sets, or momentarily dominant response tendencies, can have powerful effects in problem solving. Some tasks raise problems for human adults only because of wrong sets; under other sets there is no problem. Perhaps because of this, much of the literature on set concerns negatively transferring sets.

Simple Sets. Nearly all studies of what will here be called simple sets have used either water jars problems or anagrams. Since a num-

ber of these studies have been reviewed by Chown (1959), only studies not covered in her paper will be cited.

The standard procedure with water jars (see Chown) may induce large amounts of set; Luchins (1946) reported that 83% of experimental Ss (those given training problems) made set responses on the transfer problems, whereas only 0.6% of control Ss (no training problems) showed set. However, the amount of set with either water jars or anagrams is influenced by a number of variables. Set was increased by increases in the number of training problems (Mayzner, 1955; van de Geer, 1957), by speed instructions (van de Geer), by similarity between training and test anagrams (Maltzman, Eisman, Brooks, & Smith, 1956), and by unsolvable training problems in some cases (van de Geer). Studies employing other variables that have increased set are cited by Chown.

Most of the above findings are clear-cut, but some qualifications should be noted. Van de Geer's (1957) results were chiefly in the form of interactions among his six variables. Thus, he found that unsolvable training problems increased set only in boys and only if extinction problems were not given prior to transfer problems. Also, increasing the number of training problems increased set most clearly when extinction problems were not given. Rhine (1957) found that appropriate set (similar training and test anagrams) facilitated test performance only when training anagrams were difficult and Ss had experienced some failure. With easy anagrams and success experiences, there was no difference between groups trained under appropriate or inappropriate set.

Set was decreased by extinction problems given prior to test problems (van de Geer), by increasing the number of water jars (in training problems, test problems, or both, Benedetti, 1956), and by interpolating problems having different solutions among the training problems (Mayzner, 1955; Mayzner & Tresselt, 1956). Since distributed practice has been found to reduce set (Chown), the Mayzner, and Mayzner and Tresselt experiments are confounded because interpolating problems during training necessarily distributes practice on training problems.

The set studies demonstrate specific positive or negative transfer from one response pattern to another, the direction of transfer depending on the particular relation between training and transfer task. However, it would be expected that a series of training prob-

lems would also produce a nonspecific positively transferring effect: learning how to learn (Duncan, 1958; Harlow, 1949), or perhaps learning how to solve. Nonspecific transfer was demonstrated by Goodnow and Pettigrew (1956). Groups were first trained to respond to specific stimulus patterns in a two-choice situation, next were given random stimulus presentations (presumably to extinguish responding to patterns), and finally were tested for learning of new patterns. Such groups tended to learn new patterns more rapidly than Ss with no prior training. The authors believe that Ss without prior training have trouble because they pay too much attention to their own response patterns rather than to the stimulus patterns, and that this tendency may be a source of difficulty in a variety of problems. In any case, their results suggest the possibly powerful effects of nonspecific transfer, learning to think or learning to solve, in all kinds of problems, effects which have been recognized by only a few writers (Harlow, 1949; Underwood, 1952; Weaver & Madden, 1949). All water jar and anagram studies probably included some effects of learning to solve, in addition to specific positively or negatively transferring habits and sets.

With a different type of problem, but one which involved set in some sense, Lawson, Hillix, and Marx (1955), and Hillix, Lawson, and Marx (1956) found no effect on transfer problems of number of reinforcements during training, and little effect of similarity between training and transfer tasks. However, the problems (guessing circuits in a matrix of lights) differed widely from those usually used in set studies, and their Ss may have been able to discriminate fairly well between training and transfer tasks. In the usual set study, S has no way of knowing which is the first test problem (at least until he has solved it), a fact which probably tends to increase set.

Very few investigators have used anything other than water jars or anagrams to study simple set, so practically all information comes from two rather similar types of problems. Other problems are needed, as well as methodological work on water jars and anagrams. Frick and Guilford (1957) do not think that water jars induce a set of any considerable strength, and agree with Levitt (1956) that the problems are not a good psychometric or experimental instrument. No thorough methodological study of anagrams was found, although Wiggins (1956), in one part of his study, revealed a source of un-

controlled variation in anagrams with two solutions. He scaled such anagrams in terms of the frequency with which one or the other solution was given by naive Ss and found variation over the entire range (.50 to .99 probability of occurrence of one of the solutions). Wiggins went on to show that training, in the form of brief study of the list of words which were the infrequent solutions, produced changes from giving the frequent to giving the infrequent solution. Anagrams in which neither solution was especially predominant originally were more subject to change. This experiment suggests that in studies of set, use of double-solution test anagrams which have, initially, equally likely solutions would produce a worthwhile reduction in variability.

It is unfortunate that investigators of set almost never presented learning curves for either training or test problems. Analysis by stage of practice can reveal important information. For example, instructions to induce appropriate set may produce better performance on early, but not on late, training problems because set (or habit strength) can also be developed by solving a series of problems of the same class. Learning curves for transfer problems would reveal the locus as well as the persistence of transfer effects, e.g., groups with inappropriate set might show negative transfer on early test problems but not on later test problems because of learning how to solve. Van de Geer (1957) found that training conditions had different effects at different stages of transfer practice.

Although there are other factors that affect set (e.g., subject variables, see later), the papers already reviewed reveal that quite a lot is known about the functional relationships between a number of independent variables and simple problem solving sets. At the same time, most of the information comes from water jars and anagrams, tasks that are sometimes held to exemplify only reproductive, not productive, thinking (Maltzman, 1955). Even if one does not (as the reviewer does not) hold to this distinction between different kinds of thinking or problem solving, there is no question that much more needs to be known about set in more complex problems. Certain difficult "insight" tasks, such as the pendulum solution of the two string problem, appear to be problems only because the situation evokes strong, though labile, response tendencies that do not lead to solution. Some information about sets in more complex prob-

lems is developed in the several types of experiments reviewed in the next section.

Complex Sets: Functional Fixedness and Preavailability. All of these studies may be described as attempts to produce positive or negative transfer to a problem by procedures intended to change the order of dominance either of responses in a hierarchy, or of whole hierarchies.

Duncker's (1945) work introduced a type of complex set called functional fixedness, which may be defined as inhibition of use of an object in one function due to recent prior experience with the object's serving a different function. Chown reviews most of the functional fixedness studies that have appeared since Duncker's work. In a more recent study, van de Geer predicted that if an object were first used in an unusual function, no functional fixedness would be found when the object subsequently had to be used in a usual function (the typical order in functional fixedness studies is usual function first, unusual function second). A screwdriver and a wrench were available to loosen a screwhead bolt, or to serve as pendulum weight in the two-string problem. Although the number of Ss was small, the results appeared to confirm the prediction. The group that solved the problem last tended to avoid the object used just previously to loosen the bolt, i.e., showed functional fixedness. But the group that solved the problem first did not tend to avoid, in loosening the bolt, the object that had been used as a weight.

Functional fixedness is a complex set with negative transfer effects. What are here called preavailability studies are attempts to induce complex sets with positive transfer effects. Saugstad (1955) presented, one at a time, the various objects necessary to solve the Maier candle problem and had S list all possible functions for each object; this was called an "availability" test. On the test, 13 out of 57 Ss gave evidence that the necessary functions were available, i.e., listed functions that would later be necessary to solve the problem. All of these 13 Ss later solved the problem, whereas Saugstad reported that only 58% of those who did not indicate that the necessary functions were available solved the problem. Although the experiment is not impressive statistically, it does suggest that problem solving was influenced by the preavailability (set, dominance) of crucial responses in the hierarchy of responses associated with each component of the problem.

Staats (1957) had Ss list uses for a screwdriver and other objects, then solve the two-string problem with the screwdriver as the only object heavy enough to serve as pendulum weight. Only 7 of the 61 Ss initially indicated using a screwdriver as some sort of weight, whereas 55 Ss eventually solved the problem. Although this portion of the experiment was inconclusive, Staats did find low but significant correlations between time to solve and frequency and latency of weight responses given in a postsolution listing of screwdriver uses. He believed that these correlations between verbal (listed uses) and instrumental (problem solving) response hierarchies indicate that problem solution would have been facilitated if weight responses had been elicited in sufficient numbers prior to solution.

A different method of manipulating preavailability was used by Judson, Cofer, and Gelfand (1956). Their Ss first learned several 5-word lists, among which were included words, in various numbers and in various contexts, relevant to solution of the later-presented problem. Thus, *rope*, *swing*, and *pendulum* were presumably relevant to the two string problem; *prop*, *ceiling*, and *floor* were relevant to the Maier hatrack problem. In general, the group that learned a list containing all three key words was better than other experimental and control groups at producing pendulum solutions to the string problem, or floor-to-ceiling solutions of the hatrack problem. Not all of the many differences (there were two replications of the string problem experiment) were statistically significant, and the findings were limited to men. Women produced few solutions of the desired type to either problem.

Judson et al. also reported an experiment showing that reinforcement of one response, taken from a previously elicited chain of free associations, significantly increased the probability of occurrence of other words in the same chain. Brief mention was also made of two attempts to facilitate solution of the string problem by prior elicitation of free associations to a list of words, one of which was *rope*. In the first experiment, those who had given "swinging" associations to *rope* produced significantly more pendulum solutions than those who had not, but these results were not confirmed in the replication. However, the general trend of all their experiments tended to support their notion that set and direction in problem solving can be interpreted in terms of response hierarchies which are influenced by characteristics of the problem and by reinforcement.

In an oft-cited experiment, Maier (1930) claimed to have demonstrated that relevant past experience is not always sufficient to solve a problem. The *Ss* must also have "direction," a sort of set or connection between past experience and present problem to enable them to bring the relevant past experience to bear. Weaver and Madden (1949) and Saugstad (1957) repeated the essential parts of Maier's experiment by comparing groups given only the relevant past experience with groups given past experience plus the hint that supposedly serves as direction. Neither found that addition of direction increased number of solutions. Saugstad also experimented with the three part-tasks Maier had used to provide relevant past experience for the test task (two-pendulum problem). Solutions of the two pendulum problem increased directly from demonstration of the part-tasks (Maier's procedure), to solving them as problems themselves, to solving them when one of the three was presented in an improved version. Saugstad held that "availability of functions" was all that was necessary to solve the problem.

Weaver and Madden pointed out that Maier ignored nonspecific past experience (learning to learn?) in the form of habits of searching and exploration, habits which may be transferred directly to the present problem without the aid of direction. Nevertheless, Maier, with his concept of direction, early called attention to the fact that merely having relevant past experience is no guarantee that *S* can bring it to bear to solve a problem. This is an issue that runs through much research and discussion of problem solving in human adults; adult *Ss* "know" the correct responses, but do not have the correct set.

One other study might be classified under preavailability. Kolars (1957) used problems requiring abstraction among forms presented on a screen. A cue form that would aid solution of the problem was flashed subliminally just before presentation of the problem. The results were unclear in the first experiment, but in a second experiment there was some evidence that the subliminal cue aided problem solving.

A possible reason why the preavailability studies did not yield clear-cut results is that the various situations were usually needlessly complex, even cluttered. (To some extent this was also true of the functional fixedness studies.) For example, *S* was asked to list uses

for several irrelevant objects as well as the crucial object, or was asked to solve the problem in the presence of several irrelevant objects. No necessary purpose seems to be served by these or other ways of complicating the situation. Moreover, such complex situations probably generate a potpourri of positive and negative sets which are difficult to analyze and which may increase variability. Judson et al. (1956) seemed to be implying this criticism of over-complexity when they suggested that their experiments may have been overcontrolled.

In spite of the sometimes ambiguous results, and in spite of the small number of published studies, functional fixedness and preavailability experiments represent, in the reviewer's opinion, one of the most fruitful types of research on problem solving. If one takes the position that for human adults many problems are such that they demand responses which are low in a hierarchy, then functional fixedness and preavailability studies are seen as direct attempts to manipulate such hierarchies. With more detailed analysis of problem situations, and with refinement of methods, such studies could contribute much to knowledge of the antecedent conditions of problem solution.

Order Effects. For present purposes, experiments in which an attempt was made to influence problem solution by varying the chronological order of certain experiences are classed as studies of set. Some papers already reviewed dealt in part with effects due to order of experience (Kendler & Kendler, 1956; Maltzman, Eisman, & Brooks, 1956; Székely, 1950b; van de Geer, 1957).

Stolurrow, Hodgson, and Silva (1956) found negative transfer in airplane mechanics from both orders of presentation of school training and brief job experience. Herman and Engstrand (1957) devised two classes of problems, one depending on position of letters on the cards, the other depending on relationships in the alphabet. The results showed: positive transfer between problems of the same class, zero transfer from position to alphabet problems, negative transfer from alphabet to position problems.

Swartz (1955) did not find any effect on solution of a problem devised from playing cards by prior sorting of the cards into suits.

It was pointed out earlier that differential, and unknown, transfer effects may have been operating in a number of problem solving

experiments. The research on order of presentation suggests that in designing studies of problem solving, one should not ignore the possibility that the experimental design may permit, even reinforce, differential transfer effects.

In over-all view, this major section on training and transfer in problem solving appears as follows. In the case of problems which depend on simple sets, the effective training variables were largely the same variables, operating in much the same way, that influence transfer performance in other learning situations. No such summary statement can be made about the antecedent variables for any other types of problems, although a few kinds of complex sets had some effect. In part, research on complex problems has yielded conflicting results; more importantly, too little research has been done. Furthermore, experiments on complex problem solving are mostly of the simple two-group type; studies in which even one variable was systematically manipulated over a wide range are almost nonexistent. Systematic variation cannot, of course, be undertaken until variables are identified and dimensionalized, but little analytic work of this kind has been done in research on complex problem solving.

Variation within the Problem

In this group of studies, either conditions concurrent with the problem, or characteristics of the problem itself, were varied. The experiments are extremely heterogeneous. Because of this, no good defense can be offered for the subcategories used.

METHODS OF PRESENTING THE PROBLEM

This category includes studies in which the same problem was presented in different modes or appearances. The different modes were usually, but not always, isomorphic to each other in the sense that relationships among the elements of the problem remained the same.

Concreteness. Many problems can be presented in either symbolic or concrete (real) form, in various degrees of these extremes, in miniature scale models of the real presentation, etc. Also, degree of overtness of S's behavior, insofar as it is under the investigator's control, has been used as a method of varying concreteness.

The following studies found no effect of varying concreteness of the problem, or in some cases, of concreteness of S's behavior: Saugstad (1957) with a miniature scale model vs. the real presentation of the two pendulum problem; or Lorge, Tuckman, Aikman, Spiegel, and Moss (1955a, 1955b), who used the mined road problem at seven "levels of reality" (verbal, photographic, miniature scale model, or real presentation, or various amounts of manipulation of the scale and real versions). In Saugstad's repetition of the Maier experiment, the "direction" was a hint that was supposed to call attention to the ceiling. Saugstad thought that a miniature scale model of the actual hallway in which the two-pendulum problem usually must be constructed, would call more attention to the ceiling, but neither number of solutions nor behavior of failing Ss gave any indication that the ceiling was a special source of difficulty.

In contrast to the preceding studies, Cobb and Brenneise (1952) and Gibb (1956), found rather clear-cut effects by varying concreteness. Cobb and Brenneise reported that anchor, reach, and extension solutions of the two-string problem decreased as concreteness decreased over four steps. Pendulum solutions were little affected but were few enough so that for all types of solutions combined, percentage solutions were perfectly correlated with increasing concreteness. Gibb used three types of subtraction problems presented in three degrees of concreteness to second grade children. Both main variables were significant on most measures, and did not interact. If children are more affected by concreteness than are adults, Gibb's results would not necessarily conflict with the studies reporting no effects of concreteness. But there is no obvious way of accounting for Cobb and Brenneise's positive results. They did use what is probably more of an "insight" problem than did the studies reporting negative results, but it was only the insight (pendulum) solution that was not affected by concreteness. Their least concrete mode of presentations seems qualitatively different from the other three modes, but this would not account for all their results.

Distribution of Work and Rest. Periods of work and rest on a problem can be varied in a number of ways. Riley (1952) found no clear-cut difference between intertrial rests of 8 sec. vs. 2 min. during learning of a rote list that required S to discover, to varying degrees, the response term for each stimulus. He noted that if anything, his

results were the reverse of the hypothesis that massing of practice should produce better performance early in learning because it should facilitate discovery, whereas distribution should be better later in learning because it should facilitate fixation (Underwood, 1949, reviews the older studies from which this hypothesis was developed).

Distribution of practice had clearcut effects in Shaklee and Jones' (1953) experiment when work and rest cycles were varied prior to solution of a kind of matching-by-inference problem. Groups worked under continuous practice, under cycles of 1 min. work-30 sec. rest, or cycles of 1 min. work-4 min. rest. In a second experiment the latter cycle was changed to 1 min.-90 sec. cycles. In both experiments, the first and third cycles, i.e., continuous practice and the quite distributed cycle, did not differ in terms of percentage of solutions, but both were significantly superior of the 1 min.-30 sec. cycle. This U-shaped function between correct solutions and distribution did not occur with incorrect solutions, which increased directly with distribution of practice.

It is rather clear that distribution of practice in problem solving needs further study.

Other Methods. Each of the following studies used a unique method of presentation. Katz (1949, experiment more briefly reported in 1950) had adult Ss give sums based on the numbers 1-9; with children, the numbers were 1-5. Each number was printed on a card. The cards were presented in what might be called "degrees of disorder," e.g., cards were presented in order in a column, in an unordered column, after being shaken in a box, etc. Time to give sums, and errors, increased directly with increasing disorder of presentation both in children and in adults.

The calculus of propositions tasks (see Moore & Anderson, 1954b) were presented by Anderson (1957) as if they had from 1-4 goals or solutions, when in fact there was only one goal. The number of Ss achieving the goal decreased directly as number of stated goals increased. This result may be roughly similar to one which apparently occurs with the two-string problem. Instructions to find as many solutions as possible, vs. insistence on the pendulum solution only, seem to elicit anchor, reach, and extension, at the expense of pendulum, solutions.

Two other studies found no effects of different methods of presenting the problem. Fattu and Mech (1953b) reported no differences attributable to interrupting Ss at various stages of work on gear train problems and asking them to state verbally where the malfunction was. Hafner (1957) found no effect in fourth grade children of instructions to verbalize while working on a stencil design problem.

In general, degree of concreteness has had little effect on problem performance in adults, except in Cobb and Brenneise's (1952) experiment. Studies using other methods of varying presentation are too few or too dissimilar to summarize.

VARIATION AMONG ELEMENTS OF THE PROBLEM

These studies are also, in a way, methods of varying presentation of the problem, but in this case there was usually a real change in the problem itself, e.g., a change in the number, order, or kind of problem elements. Perhaps the Katz (1949) and Anderson (1957) experiments could just as well have been included here, as well as some experiments on simple sets. In the latter, it has been found that interpolation of various conditions among the test problems will reduce set, e.g., extinction problems (van de Geer), additional jars (Benedetti).

In Judson and Cofer's (1956) experiment Ss had to select the word that was out of place in groups of words, each group containing two ambiguous and two unambiguous words. The Ss clearly chose on the basis of the first-appearing unambiguous word; in the authors' terms, "priority of activation of a response hierarchy" significantly influenced behavior. Increasing the number of ambiguous words between the two unambiguous words increased the dominance of the first-occurring unambiguous word.

Surprisingly strong effects of spatial contiguity among elements of a problem were reported by Kay (1954). The Ss had to turn off a row of lights three feet away from a row of switches, using as a cue numbers printed in a random arrangement on a card. When a light came on, S assigned it a number from 1 to 12 (left to right) located the number on the card, and pressed the switch in line with the number. Time and error scores increased directly as the card was

first placed directly in front of the switches, then moved to midway between, finally placed directly in front of the lights. A few Ss, especially older ones, could not do the task at all if the card was anywhere beyond the midpoint, i.e., closer to the lights than to the switches. The effects of contiguity might not have been so great if Ss had performed the most difficult task (card directly in front of lights) first, then transferred to the easier tasks. Nevertheless, the results clearly suggest that intraproblem contiguity is of fundamental importance in problem solving. It is possible that contiguity among the elements of a concrete problem heavily determines the degree to which such processes as reordering and restructuring (Wertheimer, 1945) can occur.

Solley (1957) made use of the fact that the meanings of *small*, *white*, *light*, and *up* tend to be positively correlated, and opposite to *large*, *black*, *heavy*, and *down*. Different sets of discs (boxes), each incorporating one of these dimensions, were used in the disc transfer problem. In six trials through the problem there were fewer errors when boxes had to be moved in the normal light-to-heavy direction than in its opposite. The ordinary size-cue expectancy, small-to-large, produced fewer extra moves and shorter time than its opposite. Other comparisons were not significant.

In Cobb and Brenneise's experiment, anchor, reach, and extension solutions of the two-string problem decreased, pendulum solutions increased, when the investigators changed the group of objects ordinarily available to an alternative group that was more relevant to pendulum solutions.

Studies of behavioral processes in problem solving (see later) sometimes also report changes in performance due to variation among problem elements. Battig (1957) had Ss guess the letters of a word with foreknowledge only of the number of letters in the word. The particular words used were a major source of variance; both length and frequency of usage of the words were complexly related to the several response measures. Hunter (1957) used different ways of stating problems of the type: A is greater than B, C is greater than B, which is greatest. There were differences due to the ways of stating the problems, to atmosphere effects, and to type of relation used (happier-sadder, taller-shorter, etc.). Hunter's study has some similarity to earlier research (not reviewed here) on atmosphere and order effects in syllogistic reasoning.

In contrast to the experiments on methods of problem presentation, studies of variation among problem elements consistently reported at least some significant effects, occasionally powerful effects, on problem solving performance. Thus, performance on a problem may or may not be influenced by contextual variables, such as methods of presentation that do not change relationships among elements of a problem. But changes of a problem's internal structure usually influence performance, even in cases where the problem remains, in some physical sense, the same.

DIFFICULTY

All variables that significantly affect speed or frequency of solution could be said to influence the difficulty of a problem. The studies to be reviewed here are those in which some condition intended to influence difficulty was deliberately varied.

All experiments on methods of "understanding" (Corman, 1957; Crannell, 1956; Forgas & Schwartz, 1957; Hilgard et al., 1953; Hilgard et al., 1954) used several transfer tasks, some called simple, others difficult. In some cases, but not always, it appeared that different training methods produced differences only on difficult transfer problems.

Performance is rather clearly affected by deliberate increases in problem difficulty. Within limits, problem difficulty is increased by increasing: the number of stimulus items with number of response items held constant (Brush, 1956), the number of stimulus-response or total items (Brush, 1956; French, 1954), or the response availability, defined as the number of response items from which the correct response for each stimulus must be selected (Brush, 1956; Noble, 1955; Noble, 1957; Riley, 1952). These studies make an important contribution to knowledge of S-R relationships in problem solving; in particular, the response availability experiments represent direct attacks on the important dimension of response discovery. The most extensive work on response availability is that by Noble. He showed that with four stimuli, difficulty increased directly as number of available responses per stimulus increased from 4 to 10, but that there was relatively little further increase in difficulty from 10 to 14 alternatives.

Ling (1946) and John (1957) give detailed protocols of changes in behavioral processes that occur when problem difficulty is increased.

Ling used Köhler-type tool problems of increasing difficulty with young children. John developed a complex device called the PSI (Problem-Solving and Information) Apparatus (see also John & Miller, 1957) which was used with two levels of difficulty with adults. In the Goldbeck et al. (1957) study of the half-split technique, use of different levels of difficulty on their apparatus revealed that the technique was of no value on the more difficult problems until Ss were first given training on deductive skills.

As might be expected, performance usually varied as a function of problem difficulty. Noble's work shows that the function is not necessarily linear.

HINTS AND AIDS

Various hints, aids, or instructions, given *S* just before or during work on a problem, have been used to facilitate solution. Maltzman, Eisman, Brooks, and Smith (1956) found that instructions influenced solution of test anagrams regardless of the class of training anagrams or the type of instructions that *S* had been given for training anagrams. In one of Burack and Moos' (1956) experiments, three increasingly-concrete hints concerning the principle of centrifugal force were given one at a time at 2-min. intervals while *S* worked on the mechanical puzzle. After all hints had been given, five of the eight Ss had managed to solve the problem.

Experiments in which aids of various kinds were of more primary concern have been reported by Reid (1951) and Marks (1951). Reid's study was based on Duncker's (1945) notion of "explication of the goal." Several experiments were done on two problems: make triangles out of matches, and fit together pieces of wood to form a tetrahedron. Experimental groups received hints at regular intervals while working, each successive hint making the goal increasingly more explicit. The hints for control groups were not intended to explicate the goal. In general, each successive hint to experimental Ss increased the number of Ss solving the problem; eventually, significantly more experimental Ss solved in all experiments and on both problems.

Marks' Ss tried to locate errors in square root problems. Two kinds of hints were used, a list of possible sources of error, or *E*'s urging *S*, at intervals, to ask himself where a mistake could occur.

As Marks predicted, verbal urging increased both *S*'s vocalizations (naming or pointing to problem elements), and the number of solutions, but contrary to prediction, the list had no effect on number of solutions. Verbal urging yielded tetrachoric correlations of .94 with vocalizations, .82 with solutions.

All of the studies on aids found that at least some kind of aid was effective, sometimes very effective. It is curious, then, to find an occasional study reporting that *Ss* were aided if necessary, but not reporting how many *Ss* were aided or if aid had any effect.

In summary of this major section on variation of conditions during the solving of a problem, it may be noted that almost all variables studied have influenced performance. The major exception is the class of diverse procedures called methods of problem presentation which, except perhaps for concreteness, yielded either conflicting results or too few results with any one method to warrant a conclusion.

The studies reviewed in this section illustrate a weakness that runs through the whole area of problem solving research, viz., the heterogeneity of problems and techniques employed. About 100 empirical studies, several of them including more than one experiment, are covered in this review. In nearly half of these studies the problem used was devised by the authors and has not yet been used by anyone else; even a brief description of each of these problems would have added materially to the length of this paper. This diversity is a major reason why the area of problem solving seems so chaotic, and is a serious obstacle to systematic progress. A few authors stated the advantages that their new problems were presumed to have, occasionally in separate publications (Marx, Goldbeck & Bernstein, 1956; Moore & Anderson, 1954b), but most did not. Problem solving research would be improved if more efforts were made to meet the standards for problems set by Ray (1955).

Individual vs. Group Problem Solving

Several carefully done experiments in the recent literature bear on the question of whether groups solve problems better than do individuals. Taylor and Faust (1952), and Lorge, Tuckman, Aikman, Spiegel, and Moss (1955a, 1955b, 1956) found groups superior on at

least some response measures. Taylor and Faust compared individuals, groups of two, and groups of four, on the game of "Twenty Questions." All Ss were instructed that number of questions was the important score. On number of questions and on time, twos and fours did not differ, but both were significantly better than individuals. Failures decreased directly from individuals to twos to fours (all differences significant). On an efficiency measure (man-minutes: number of persons $\times$ time), individuals were better than twos, twos were better than fours. The authors could also have shown that individuals took by no means twice as many questions as twos or four times as many as fours. Practice effects over days did not differ as a function of the three conditions.

The first two studies by Lorge et al. (1955a, 1955b) were cited earlier in connection with their seven methods of presentation of the mined road problem. Individuals were compared with groups of five under all methods. For scoring, a content analysis was made of S's written solutions and crucial aspects of the solution were weighted. There were highly significant differences in favor of groups over individuals on this "quality-of-solution" measure under all methods of presentation, with no interaction. Groups asked more questions than individuals, suggesting that group superiority was in part due to obtaining more information. In another experiment (Lorge et al., 1956), only the real presentation of the problem was used. Group superiority was again evident. It was also shown that in their written reports, groups tended to underestimate the quality of their actual solutions (as measured by reliable observers). In part, individuals tended to overestimate their solutions. In these several experiments, Lorge et al. did not report an efficiency measure; groups were better in over-all quality of solution, but were almost certainly not five times better either in over-all quality or in any one component of solution.

McCurdy and Lambert (1952), Moore and Anderson (1954a), Marquart (1955), and, perhaps, Comrey and Staats (1955), found no evidence for group superiority. The McCurdy and Lambert problem required turning six switches. Working individually, S turned all switches; working in groups of three, each S turned two switches. Groups were no better than individuals, and leaderless groups were no different from groups in which one S gave directions that the others had to follow.

Moore and Anderson first matched groups of three Ss with individual Ss on knowledge of the calculus of propositions tasks. Over a 10-day period of solving problems, individuals did not differ significantly from groups on: number of problems solved, mean steps taken on problems, mean time on solved problems, mean errors, or on two measures of repetitiousness of response. On a man-hour basis, individuals were almost three times as efficient as groups. Moore and Anderson had forced groups to agree on steps in solving, so one member would not dominate; thus, they noted that groups had to work as groups, a responsibility not saddled onto individuals.

Marquart (1955) repeated and expanded the oft-cited Shaw (1932) study, using eight problems of various kinds. All Ss worked on all problems, both as individuals and as members of groups of three. By Shaw's method, involving comparison of total solutions to total possible solutions, groups were superior. But since this method does not indicate whether a group solution was merely due to the best member, Marquart combined individuals into "groups" of three. By this method, groups working as groups were no better than groups working as individuals. Marquart also used her method to reanalyze Shaw's data, and found little difference between Shaw's groups and individuals (see pp. 537-549).

The Ss solved crossword puzzles in Comrey and Staats' (1955) study, first solving individually, then in pairs where one S had the vertical code, the other the horizontal code. It was shown that 82% of the variance on the group task could be predicted from a linear combination of perfectly reliable high and low individual scores.

The results of the preceding group of experiments can be fairly easily summarized. On "over-all" types of measures, groups have been superior to individuals on a few problems, but not on most problems. But where efficiency measures were reported, and also probably where they were not reported, individuals were superior.

Although theories of problem solving will be taken up later, Lorge and Solomon's (1955) paper is of interest here since it deals with two models of group problem solving. Working with Shaw's (1932) data, Lorge and Solomon noted that some groups solved all the problems, some solved none. This suggested that group superiority was due to the abilities of members of the group, rather than to interpersonal interaction. So two ability models were proposed:

(A) group superiority is a function only of the ability of one or more of its members to solve the problem without regard to acceptance or rejection of members' suggestions, (B) group superiority is a function only of the pooled abilities of its members. Pooled abilities can produce solutions even though no member of the group can solve alone. Model B implies that any problem may be composed of, and solved in, two or more stages, so it reduces to Model A for one-stage problems. Model A was found to be tenable for two of Shaw's three problems. It was also shown that Model A can be modified for stage-wise solutions. From applying Model B, the authors concluded that Shaw's data suggest, not personal interaction, but pooled abilities in two-stage problems, i.e., Model B. Perhaps Marquart's (1955) method of combining individuals into "groups" to compare with actual groups, and her finding that these two types of groups did not differ, is a statistical pooling of abilities which produces solutions even without face-to-face contact.

Problem Solving Processes

There seems to be more concern with behavioral processes, as presumably different from products, in the field of thinking and problem solving than in any other area of learning or performance. As compared to the literature before 1946, recent investigators tend more and more to report only products, e.g., so many Ss solved the problem, so many did not. Even so, perhaps half of all papers cited in this review have had something or other to say about processes. Obviously, only major studies of processes can be summarized here, and these only very briefly. Studies of, or discussions about, problem solving processes are often long and extremely detailed.

"Processes" can mean almost anything: insight vs. trial and error, response variability, flexibility vs. rigidity, methods of attack, basic processes such as perception, memory, intelligence, learning, etc. Other so-called processes are sometimes named and described in terms of the characteristics of the particular problems used in a study. This diversity precludes any close comparison of the results of different studies. Furthermore, processes are sometimes studied merely by giving a single group of Ss a problem and describing the Ss' behavior in verbal or frequency distribution form; there may be

little attempt to quantify processes or to vary conditions. Some of the distinction between process and product would disappear if more efforts were made to determine functional relationships between dimensionalized processes and systematically varied conditions.

Bloom and Broder's (1950) remedial work with failing college students was based on observations of successful and unsuccessful problem solvers, i.e., students who did well or poorly on problem solving types of examinations. Detailed descriptions of differences in problem solving behavior, and in personality, between good and poor solvers were reported. The Ss' responses were classified under: understanding the nature of the problem, understanding the ideas contained in the problem, general approach to the solution of problems, and attitude toward problem solving. All of these classes revealed differences between good and poor solvers. The authors also noted that good and poor solvers differed not so much in having relevant information, but in applying it to a problem. McNemar (1955) reported a somewhat similar finding. Bloom and Broder claimed that problems have a figure-ground organization in that some elements of a problem stand out much more than others, and that some elements, not necessarily figural ones, furnish starting points much more than others. It seems likely that such figural elements of all kinds are prime inducers of sets.

In Buswell's (1956) very extensive study, over 500 Ss were observed while working on various mathematical problems, while attempting to discover and transfer generalizations (this portion of the study was cited earlier), and while selecting from cards the steps and methods they wanted to use in solving a problem. Buswell reported great individual differences, and trial and error rather than systematic approaches. In no case were as many as 20% of the group represented by any one pattern of thinking; the evidence gave no support to any notion that problem solving must follow precise recipes.

Earlier it was mentioned that John (1957) found rather consistent differences between those trained in natural sciences and those trained in other disciplines on his PSI problem. Actually, John tested six groups, varying in kind and amount of educational background, on two levels of difficulty of the problem. Eight work vari-

ables, four information variables, and nine approach variables were studied and intercorrelated, and changes in these patterns of behavior from the simpler to the more difficult problem were reported. These data cannot be summarized here, nor can John's over-all description of the problem solving process. Some of the points that were emphasized were that past training and experience brought about habituation of an individual to certain kinds of conceptual and organizational processes which were consistently displayed, that some aspects of personality were reflected in the problem solving process, and that present level (not type) of academic training did not appear to change parameters of effectiveness on the problems to any great extent.

Goldner (1957) studied "whole-part approach" and "flexibility-rigidity" on six problems. Although there were the usual individual differences, intra-individual consistency was fairly high from problem to problem on the whole-part variable. Flexibility-rigidity was also fairly consistent on similar tasks, but not on tasks that differed in structure. The two dimensions were separate processes in less structured problems, but were closely related in more structured problems.

Practice effects, which seem ubiquitous in other areas of performance, have not always been found in problem solving. One example (there are a few others) occurred in Bendig's (1953, 1957) work on patterns of behavior in solving twenty questions problems. Bendig's interest was in the information transmitted by questions and used by Ss. Although there were changes over problems in some of the information measures, and other significant effects, there was no learning, at least by Bendig's method of measurement, over problems in either study. These results concerning practice effects conflict to some extent with those of Taylor and Faust (1952), but Bendig did not use the twenty questions game in the usual way, and Taylor and Faust's work was conducted over a much longer series of problems.

Two other major studies of problem solving processes are those by Moraes (1954), and Süllwold (1954). Part of Moraes' study was cited earlier in other connections; the major part of the work is the detailed protocols of thinking processes obtained by comparing children who were good vs. those who were poor at arithmetic reasoning.

Süllwold claimed that problem solution has two phases, one of sudden insight, the other where progress is slow, and that individual differences in exhibiting these phases were consistent from problem to problem. However, van de Geer rather thoroughly disputes Süllwold's claims.

The preceding studies do not exhaust all that has been said in the recent literature on problem solving processes (see Chown). Most of the theoretical articles to be reviewed later, as well as all of the books and a good many of the empirical investigations that were cited in other connections have included something about processes. Among empirical studies extensive data or discussion relevant to processes appear in Ling (1946), Székely (1947), and Weaver and Madden (1949).

In the reviewer's opinion, it would be preferable to devote more effort to determination of functional relationships between environmental or task variables and performance or product, rather than to problem solving processes. In oversimplified terms, determination of what the simple laws are must precede attempts to determine why and how the laws operate. At the same time, research on processes would make a greater contribution if efforts were made to develop some sort of rough classification of behavior patterns on which investigators could agree and which would be used in more than one study. Possible starting points would be Bloom and Broder's (1950) checklist, Guilford's (1956) factors, etc. At present, the chief weakness of research on behavior patterns in problem solving is that the research area itself is so unpatterned.

Theory

It is encouraging to find that in an area as unintegrated as is research on problem solving, there are a number of good theoretical beginnings. The most thoroughgoing attempt in the recent literature to develop and test a theory of problem solving is that by Maltzman and his associates. In the major theoretical paper (Maltzman, 1955), the idea of the habit family hierarchy, derived primarily from Hull, is used. The divergent, trial and error mechanism (one stimulus leading to a hierarchy of responses in which the correct response has low initial strength), and the convergent, discrimination learn-

ing mechanism (one response is led to by a hierarchy of stimuli in which the correct stimulus is initially low in the hierarchy), are combined to assume a compound, temporal hierarchy. Reinforcement or extinction of individual members of a hierarchy are assumed to generalize to other members. Changes in order of dominance in a hierarchy, or among hierarchies, may be produced by extinction of dominant incorrect responses or response families, by increasing the reaction potential of the correct response through mediated generalization from other reinforced members, or by elicitation of fractional anticipatory goal responses. Concerning extinction of dominant incorrect responses, Maltzman pointed out that spontaneous recovery may occur; thus, interfering responses may recur repeatedly while S is working on a problem. (An example of this apparently occurred in the problem used by Kay, 1954.) Mediated generalization, a basic notion in the theory, is said to be accomplished primarily by linguistic responses. Fractional anticipatory goal responses are used to interpret set; they are responses that are evoked by instructions, hints, etc. This is a valuable suggestion; although sets of all kinds play an important role in many types of performance, they have received little attention from learning theorists.

Failure to solve a problem, inability to overcome wrong set, and similar phenomena, can be accounted for by Maltzman's theory. He points out that if the correct response is low in the hierarchy, generalized inhibition from repeated unsuccessful occurrences of the dominant incorrect response may reduce reaction potential of the correct response below the threshold. Also, high irrelevant drive, such as anxiety, will not only produce competing responses, but by increasing the total drive will multiply by all habit strengths, thereby increasing the advantage of a dominant incorrect response over a weaker correct response. As noted earlier, the prediction concerning irrelevant drive has received some confirmation in simple set problems (Chown, 1959; Mayzner & Tresselt, 1956). (Prediction of the effect of irrelevant drive on other problems is difficult; a dominant incorrect set should be increased in strength, but the simultaneous occurrence of competing responses might facilitate solution by increasing response variability.) Other predictions from Maltzman's theory, and some expansions of the theory, appear in

his several empirical studies, cited earlier. The theory does not come to close grips with such tasks as the two-string problem, but it is still one of the most fruitful theories yet offered in problem solving.

Borrowing from Maltzman and others, Cofer and his associates (Cofer, 1957; Judson & Cofer, 1956; Judson, Cofer, & Gelfand, 1956) emphasize particularly the role of verbal responses as mediators in response hierarchies. The several experiments (reported in the two 1956 papers and cited earlier) deal either with variables that produce changes of dominance in verbal response hierarchies, or with the effects of such changes on problem solution. The former type of experiment was quite successful; the latter was fairly successful. Although Cofer deals mostly with hierarchies among verbal responses, as does Maltzman, there are also, of course, hierarchies among instrumental responses. Staats (1957), borrowing from Osgood (1953), developed his experiment on the basis of possible relationships between verbal and instrumental hierarchies associated with the same stimulus object.

A phenomenological theory of problem solving has been presented in some detail by van de Geer (1957). Briefly, different aspects of the same object may appear in perception; therefore, situations vary in degree of "transparency." In thinking, other aspects of the situation must be explicated, thereby reducing the nontransparency of the situation. In connection with this theory, van de Geer makes a worthwhile effort to classify problems. His major categories are three "points of view" toward problems: in what way does *S* try to solve, what is the nature of the difficulty of the problem, what is the nature of the initial and the goal situations. Each point of view provides, to some extent, a classification of problems. For example, the distinction sometimes made between insight problems and trial and error problems appears, in other terms, under "nature of the difficulty."

Saying that a phenomenological theory does not lead directly to a program for experimental research, van de Geer goes on to present an axiomatic approach to problem solving, based on game theory and information theory. He shows how the model handles each of the types of problem listed under his "nature of the difficulty" category. In this connection van de Geer claims that *S*'s intelligence and "thinking-out capacity" determine how difficult *S* will find a prob-

lem to be, and therefore whether *S* will show insight or trial and error. This attempt to reduce insight and trial and error to a single underlying principle has some similarity to Galanter and Gerstenhaber's (1956) relating of the two patterns of behavior.

Van de Geer's point that phenomenological theory does not easily generate experimentally testable hypotheses also applies to some other "theories," descriptions of processes, lists of steps toward solution, etc., which have been offered in the area of thinking and problem solving. In sharp contrast, Underwood (1952) presents a combination of theory and orientation toward research that directly suggests manipulatable variables. To begin with, thinking, including both concept formation and problem solving, is said to be the learning or the recognizing (discovering?) of perceptual or functional relationships among stimuli. Stimuli may include objects, symbols, or other relationships (as in syllogisms). The basic assumption is that for the perception of relationships among stimuli to occur, the appropriate responses to those stimuli must be contiguous. The reviewer would interpret this assumption to mean that when presented separately, S_1 and S_2 lead, or can be made to lead, to the same R_1 . Since both stimuli lead to the same response, there is a relationship between them which, however, will not necessarily be perceived unless they are presented in such a way that "both" R_1 s occur contiguously. A mediational mechanism could also be included: the first R_1 produces stimuli, traces of which overlap with occurrence of the second R_1 .

Whether or not this is a correct interpretation of Underwood's basic assumption, it is clear that manipulatable variables in thinking are those factors that increase or decrease response contiguity. Underwood mentions such factors as mode of presentation of stimuli, number and similarity of stimuli, several kinds of biases, and memory. He points out the importance of response hierarchies, and indicates that the theory leads to a number of predictions. One of the predictions is that massed practice, by preventing forgetting, may be better than distributed practice in thinking. The literature bearing on this point is conflicting. Although Underwood's theory is not as easily applied to some types of complex problems as it is to other types or to concept formation, it is more directly tied to a

basic process (contiguity) than is any other theory, and is one of the best single sources for research hypotheses.

Except for van de Geer's phenomenological theory, all of the theories discussed so far are S-R behavioristic types. Unfortunately, much less has been done to expand Gestalt theory. Distinctions between productive and reproductive thinking, discussions of the role of insight, and experiments on functional fixedness, explication of the goal, and water jars set, all stem more or less from Gestalt origins. Humphrey (1951), and van de Geer (1957) indicate strengths and weaknesses of Gestalt theory, and Saugstad (1957) rejects it. But only Helson and Helson (1946) have made a serious attempt to generalize the theory to new situations. Their approach is to show that configurational principles also apply to abstract, symbolic problems, as well as to Wertheimer's (1945) geometric tasks. They go through the steps needed to solve a mathematical problem by analysis of the whole (equation) into natural parts related to it, rather than by trial and error or by use of high-level mathematical knowledge. It is shown that reorienting the equation by use of new symbols aids solution, and this reorienting is said to be the same process that operates in geometrical problems or in perceiving hidden figures. (Others have directly attempted to relate reorganization in certain problems to scores on hidden figures tests: see Chown, 1959; Frick & Guilford, 1957.) Helson and Helson consider that substitution of symbols or of new symbols is the distinguishing mark of abstract thinking, and that it is frequently desirable to replace concrete features with symbols since symbols are easier to manipulate and also tend to suggest new combinations. This point may have some relation to the "concreteness" studies reported earlier; if Ss do tend to replace concrete features with symbols, the failure of most of the concreteness studies to find any difference among various perceptual or symbolic modes of problem presentation would be understandable. Indeed, Helson and Helson conclude that no sharp line can be drawn between concrete and symbolic procedures; most individuals use both in actual thinking.

Gestalt theory, with its emphasis on reorientation within a problem, also bears some relationships to studies of "variation among problem elements," and to studies concerned with "methods of understanding." It is possible that Ss could be trained in reorienting

as a method of understanding, and that such a skill would transfer to a wide variety of problems.

Other theoretical contributions may be mentioned briefly. Flavell and Draguns (1957) hold that both thought and perception undergo a very brief but important microdevelopment. Their suggestions deal with a matter to which too little attention has been paid, viz., the sets that are instantaneously induced upon initial perception of a problem. Stolurow, Bergum, Hodgson, and Silva (1955) present a probabilistic model of trouble shooting. The probability that each of several defects may be causing malfunction in airplane engines and the time to repair each defect are combined in a ratio to indicate which order of checking defects one should follow for most efficient repair. It seems likely that the same sort of model could be worked out for other complex apparatus problems such as Fattu, Mech, and Kapos' (1954) gear train. Humphrey (1951), Johnson (1955), and Weaver and Madden (1949) all make several points relevant to the development of problem solving theory. Mayzner (1955) first develops predictions from theories of Hull, Werner, and an earlier theory of Underwood, then shows how each theory fared in comparison to his data.

Several of the prototheories reviewed here seem promising. However, they have not yet been directly followed up by much experimentation, and those who do experimental work have made little effort to relate their results to what theory is available. This lack of *rapprochement* between existing theory and existing data is another one of the reasons why the area of problem solving shows lack of integration.

Some further points may be made with regard to theory in problem solving. First, problem solving in human adults is to a considerable extent a matter of transferring past-learned skills and responses to the immediate problem situation. In one way or another this fairly obvious point has been implied by many investigators, even those who hold to a distinction between productive and reproductive thinking. Yet relatively little use has been made of existing transfer theory or data. For example, many problems can be interpreted as negative transfer situations. Some of the variables of which negative transfer is a function are known from studies of other types of human learning and performance, both verbal and

motor. These studies provide many suggestions for research, and to some extent for theory, in problem solving. Negative transfer is merely an example. A thoroughgoing transfer approach to problem solving could also make use of much that is known about positive transfer, including the possibility of "learning to think" (Harlow, 1949; Underwood, 1952; Weaver & Madden, 1949).

Second, except for Underwood's (1952) paper, and perhaps a few suggestions by Bruner, Goodnow, and Austin (1956), almost nothing has been done to relate, theoretically or experimentally, the area of problem solving to the large literature on concept formation. Yet the initial discovery of relevant among irrelevant dimensions in concept formation is probably not basically different from discovery of the correct solution in problem solving. Both Riley (1952) in problem solving, and Richardson and Bergum (1954) in concept formation, have recognized separate discovery and fixation phases in performance. The discrete S-R problems used by several investigators (e.g., Brush, 1956; French, 1954; Marx et al., 1956; Noble, 1955; Ray, 1957; Riley, 1952) can probably be modified to vary continuously from "pure" concept formation to "pure" problem solving tasks.

Finally, despite what has been said above concerning theory, the reviewer's position is the same as that of Ray (1955) and Underwood (1952). These authors emphasize that although theoretical developments are not necessarily unwelcome, the basic need in problem solving is experimental determination of the functional relationships between dimensionalized independent variables and problem solving performance.

Conclusions

The following conclusions are suggested. Problem solving in human adults is a name for a diverse class of performances which differs, if it differs at all, only in degree from other classes of learning and performance, the degree of difference depending upon the extent to which problem solving demands location or integration of previously learned responses. Problem solving performance varied most clearly as a function of simple sets, of a few kinds of complex sets, of changes in the relationships among elements of a problem, of

level of problem difficulty, of aids toward solution, and of certain characteristics of the subject, especially sex, age, and reasoning ability. The variables that influence simple sets were largely those that affect performance, and that affect performance in similar ways, in other situations. Individual differences in problem solving proficiency appeared to be relatively stable.

Problem solving was usually, though not always, unaffected by differences in the degree of concreteness or abstractness of versions of the same problem. Other variables and conditions either yielded conflicting results, or more commonly, were employed in too few studies to warrant a conclusion.

Groups produced more or better solutions to some problems than did individuals; on most problems there was no difference. Individuals were superior to groups on measures of efficiency. Research on problem solving processes revealed very diverse patterns of behavior. Problem solving theories that show some promise are beginning to be developed.

The field of problem solving is poorly integrated. The reasons for this seem to be the use of a great variety of tasks to provide problems, the frequent use of unanalyzed and nondimensionalized variables, the lack of an agreed-upon taxonomy of behavioral processes, and to some extent the failure to relate data to other data or to theory. Problem solving particularly needs research to determine the simple laws between dimensionalized independent variables and performance.

References

- ANDERSON, S. B., Problem solving in multiple-goal situations. *J. exp. Psychol.*, 1957, 54, 297-303.
- BATTIG, W. F., Some factors affecting performance on a word-formation problem. *J. exp. Psychol.*, 1957, 54, 96-104.
- BEHRENS, B. P., & MILES, G. H., A test of "tendency to analyze" for use with college men. *Proc. Iowa Acad. Sci.*, 1957, 64, 508-513.
- BENDIG, A. W., Twenty Questions: An information analysis. *J. exp. Psychol.*, 1953, 46, 345-348.
- BENDIG, A. W., Practice effects in "Twenty Questions." *J. gen. Psychol.*, 1957, 56, 261-268.
- BENEDETTI, D. T., A situational determiner of the Einstellung-effect. *J. gen. Psychol.*, 1956, 54, 271-278.

- BLOOM, B. S., & BRODER, L. J., *Problem-solving Processes of College Students*. Univer. Chicago Press, 1950.
- BRUNER, J. S., GOODNOW, J. J., & AUSTIN, G. A., *A Study of Thinking*. New York: Wiley, 1956.
- BRUSH, F. R., Stimulus uncertainty, response uncertainty, and problem solving. *Canad. J. Psychol.*, 1956, 10, 239-247.
- BURACK, B., & MOOS, D., Effect of knowing the principle basic to solution of a problem. *J. educ. Res.*, 1956, 50, 203-208.
- BUSWELL, G. T., Patterns of thinking in solving problems. *Univer. Calif. Publ. Educ.*, 1956, 12, 63-148.
- CHOWN, S. M., Rigidity—a flexible concept. *Psychol. Bull.*, 1959, 56, 195-223.
- COBB, H. V., & BRENNEISE, S. H., Solutions of the Meier string problem as a function of the method of problem presentation. *Proc. S. Dak. Acad. Sci.*, 1952, 31, 138-142.
- COFER, C. N., Reasoning as an associative process: III. The role of verbal responses in problem solving. *J. gen. Psychol.*, 1957, 57, 55-58.
- COMREY, A. L., & STAATS, C. K., Group performance in a cognitive task. *J. appl. Psychol.*, 1955, 39, 354-356.
- CORMAN, B. R., The effect of varying amounts and kinds of information as guidance in problem solving. *Psychol. Monogr.*, 1957, 71 (2, Whole No. 431).
- CRAIG, R. C., Directed versus independent discovery of established relations. *J. educ. Psychol.*, 1956, 47, 223-234.
- CRANNELL, C. W., Transfer in problem solution as related to the type of training. *J. gen. Psychol.*, 1956, 54, 3-14.
- DUNCAN, C. P., Transfer after training with single versus multiple tasks. *J. exp. Psychol.*, 1958, 55, 63-72.
- DUNCKER, K., On problem-solving. *Psychol. Monogr.*, 1945, 58 (5), (Whole No. 270).
- ENGLEHARD, MARY DOMINIC, *An Experimental Study of Arithmetic Problem-solving Ability of Sixth Grade Girls*. Washington, D.C.: Catholic Univer. Amer. Press, 1955.
- FATTU, N. A., & MECH, E. V., The effect of set on performance in a "trouble shooting" situation. *J. appl. Psychol.*, 1953, 37, 214-217. (a)
- FATTU, N. A., & MECH, E. V., Interruption: Its effect upon performance in a "trouble shooting" situation. *J. Psychol.*, 1953, 36, 153-163. (b)
- FATTU, N. A., MECH, E. V., & KAPOs, E., Some statistical relationships between selected response dimensions and problem-solving proficiency. *Psychol. Monogr.*, 1954, 68 (6, Whole No. 377).
- FLAVELL, J. H., & DRAGUNS, J., A microgenetic approach to perception and thought. *Psychol. Bull.*, 1957, 54, 197-217.
- FORGUS, R. H., The Einstellung effect and variability of behavior. *J. gen. Psychol.*, 1957, 56, 213-218.

- FORGUS, R. H., & SCHWARTZ, R. J., Efficient retention and transfer as affected by learning method. *J. Psychol.*, 1957, 43, 135-139.
- FRENCH, R. S., The effect of instruction on the length-difficulty relationship for a task involving sequential dependency. *J. exp. Psychol.*, 1954, 48, 89-97.
- FRICK, J. W., & GUILFORD, J. P., An analysis of a form of the water jar test. *Amer. J. Psychol.*, 1957, 70, 427-431.
- GALANTER, E., & GERSTENHABER, M., On thought: The extrinsic theory. *Psychol. Rev.*, 1956, 63, 218-227.
- GIBB, E. G., Children's thinking in the process of subtraction. *J. exp. Educ.*, 1956, 25, 71-80.
- GOLDBECK, R. A., BERNSTEIN, B. B., HILLIX, W. A., & MARX, M. H., Application of the half-split technique to problem-solving tasks. *J. exp. Psychol.*, 1957, 53, 330-338.
- GOLDNER, R. H., Individual differences in whole-part approach and flexibility-rigidity in problem solving. *Psychol. Monogr.*, 1957, 71 (21, Whole No. 450).
- GOODNOW, J. J., & PETTIGREW, T. F., Some sources of difficulty in solving simple problems. *J. exp. Psychol.*, 1956, 51, 385-392.
- GUILFORD, J. P., The structure of intellect. *Psychol. Bull.*, 1956, 53, 267-293.
- HAFNER, A. J., Influence of verbalization on problem solving. *Psychol. Rep.*, 1957, 3, 360.
- HARLOW, H. F., The formation of learning sets. *Psychol. Rev.*, 1949, 56, 51-65.
- HELSON, H., & HELSON, H. B., Some common features of concrete and abstract thinking. *Amer. J. Psychol.*, 1946, 59, 468-472.
- HERMAN, D. T., & ENGSTRAND, R., Order effect in problem solving. *Psychol. Rep.*, 1957, 3, 623-626.
- HILGARD, E. R., EDGREN, R. D., & IRVINE, R. P., Errors in transfer following learning with understanding: Further studies with Katona's card-trick experiments. *J. exp. Psychol.*, 1954, 47, 457-464.
- HILGARD, E. R., IRVINE, R. P., & WHIPPLE, J. E., Rote memorization, understanding, and transfer: An extension of Katona's card-trick experiments. *J. exp. Psychol.*, 1953, 46, 288-292.
- HILLIX, W. A., LAWSON, R., & MARX, M. H., Reinforcement of components and of similar patterns as factors determining hypothesis selection in problem solving. *J. gen. Psychol.*, 1956, 54, 39-43.
- HOVLAND, C. I., & KENDLER, H. H., The New York University conference on human problem solving. *Amer. Psychol.*, 1955, 10, 64-68.
- WUMPHREY, G., *Thinking: An Introduction to Its Experimental Psychology*. New York: Wiley, 1951.

- HUNTER, I. M. L., The solving of three-term series problems. *Brit. J. Psychol.*, 1957, 48, 286-298.
- JOHN, E. R., Contributions to the study of the problem-solving process. *Psychol. Monogr.*, 1957, 71 (18, Whole No. 447).
- JOHN, E. R., & MILLER, J. G., The acquisition and application of information in the problem-solving process. *Behav. Sci.*, 1957, 2, 291-300.
- JOHNSON, D. M., Problem solving and symbolic processes. *Annu. Rev. Psychol.*, 1950, 1, 297-310.
- JOHNSON, D. M., *The Psychology of Thought and Judgment*. New York: Harper, 1955.
- JUDSON, A. J., & COFER, C. N., Reasoning as an associative process: I. "Direction" in a simple verbal problem. *Psychol. Rep.*, 1956, 2, 469-476.
- JUDSON, A. J., COFER, C. N., & GELFAND, S., Reasoning as an associative process: II. "Direction" in problem solving as a function of prior reinforcement of relevant responses. *Psychol. Rep.*, 1956, 2, 501-507.
- KATONA, G., *Organizing and Memorizing*. New York: Columbia Univer. Press, 1940.
- KATZ, D., Blendungsphänomene und konnektive Hemmung bei Denkprozessen. *Theoria*, 1949, 15, 141-154. More briefly reported as: Connective inhibitions during thought processes. In M. L. Reymert (Ed.), *Feelings and Emotions: The Mooseheart Symposium*. New York: McGraw-Hill, 1950.
- KAY, H., The effects of position in a display upon problem solving. *Quart. J. Psychol.*, 1954, 6, 155-169.
- KENDLER, H. H., & KENDLER, T. S., Inferential behavior in preschool children. *J. exp. Psychol.*, 1956, 51, 311-314.
- KLIEBHAN, MARY CAMILLE, *An Experimental Study of Arithmetic Problem-Solving Ability of Sixth Grade Boys*. Washington, D.C.: Catholic Univer. America Press, 1955.
- KOLERS, P. A., Subliminal stimulation in problem-solving. *Amer. J. Psychol.*, 1957, 70, 437-441.
- KOYANAGI, K., An experimental study on relations between the intellectual cognition and the intellectual activity in children. *Tohoku psychol. Fol.*, 1953, 13, 100-113.
- LAWSON, R., HILLIN, W. A., & MARX, M. H., Hypothesis selection as a function of previous reinforcement of components of the hypothesis. *J. gen. Psychol.*, 1955, 53, 299-303.
- LEVITT, E. E., The water-jar Einstellung test as a measure of rigidity. *Psychol. Bull.*, 1956, 53, 347-370.
- LING, B. C., The solving of problem-situations by the preschool child. *J. genet. Psychol.*, 1946, 68, 3-28.
- LORGE, I., & SOLOMON, H., Two models of group behavior in the solution of Eureka-type problems. *Psychometrika*, 1955, 20, 139-148.

- LORGE, I., TUCKMAN, J., AIKMAN, L., SPIEGEL, J., & MOSS, G., Solutions by teams and by individuals to a field problem at different levels of reality. *J. educ. Psychol.*, 1955, 46, 17-24. (a)
- LORGE, I., TUCKMAN, J., AIKMAN, L., SPIEGEL, J., & MOSS, G., Problem-solving by teams and by individuals in a field setting. *J. educ. Psychol.*, 1955, 46, 160-166. (b)
- LORGE, I., TUCKMAN, J., AIKMAN, L., SPIEGEL, J., & MOSS, G., The adequacy of written reports in problem solving by teams and by individuals. *J. soc. Psychol.*, 1956, 43, 65-74.
- LUCHINS, A. S., Classroom experiments on mental set. *Amer. J. Psychol.*, 1946, 59, 295-298.
- MAIER, N. R. F., Reasoning in humans: I. On direction. *J. comp. Psychol.*, 1930, 10, 115-143.
- MALTZMAN, I., Thinking: From a behavioristic point of view. *Psychol. Rev.*, 1955, 62, 275-286.
- MALTZMAN, I., EISMAN, E., & BROOKS, L. O., Some relationships between methods of instruction, personality variables, and problem-solving behavior. *J. educ. Psychol.*, 1956, 47, 71-78.
- MALTZMAN, I., EISMAN, E., BROOKS, L. O., & SMITH, W. M., Task instructions for anagrams following different task instructions and training. *J. exp. Psychol.*, 1956, 51, 418-420.
- MARKS, M. R., Problem solving as a function of the situation. *J. exp. Psychol.*, 1951, 41, 74-80.
- MARQUART, D. I., Group problem solving. *J. soc. Psychol.*, 1955, 41, 103-113.
- MARX, M. H., GOLDBECK, R. A., & BERNSTEIN, B. B., An apparatus for investigating the methods humans use in solving complex problems. *Amer. J. Psychol.*, 1956, 69, 462-465.
- MAYZNER, M. S., The effects of the competition of various strengths of sets in problem-solution. *Psychol. Newsltr., NYU*, 1955, 6, 134-155.
- MAYZNER, M. S., & TRESSLIT, M. E., The effect of the competition and generalization of sets with respect to manifest anxiety. *J. gen. Psychol.*, 1956, 55, 241-247.
- MCCURDY, H. G., & LAMBERT, W. E., The efficiency of small human groups in the solution of problems requiring genuine co-operation. *J. Pers.*, 1952, 20, 478-494.
- MCNEMAR, O. W., An attempt to differentiate between individuals with high and low reasoning ability. *Amer. J. Psychol.*, 1955, 68, 20-36.
- MILTON, G. A., The effects of sex-role identification upon problem-solving skill. *J. abnorm. soc. Psychol.*, 1957, 55, 208-212.
- MOORE, O. K., & ANDERSON, S. B., Search behavior in individual and group problem solving. *Amer. sociol. Rev.*, 1954, 19, 702-714. (a)
- MOORE, O. K., & ANDERSON, S. B., Modern logic and tasks for experiments on problem solving behavior. *J. Psychol.*, 1954, 38, 151-160. (b)

- MORAES, A. M. DE M., *Recherche Psychopédagogique sur la Solution des Problèmes d'Arithmétique*. Louvain: Nauwelaerts, 1954.
- NOBLE, C. E., Compound trial-and-error learning as a function of response availability (N_R). *J. exp. Psychol.*, 1955, 49, 93-96.
- NOBLE, C. E., Supplementary report: Response availability in human trial-and-error learning. *J. exp. Psychol.*, 1957, 53, 359-360.
- OSGOOD, C. E., *Method and Theory in Experimental Psychology*. New York: Oxford, 1953.
- RAY, W. S., Complex tasks for use in human problem-solving research. *Psychol. Bull.*, 1955, 52, 134-149.
- RAY, W. S., Verbal compared with manipulative solution of an apparatus-problem. *Amer. J. Psychol.*, 1957, 70, 289-290.
- REID, J. W., An experimental study of "analysis of the goal" in problem-solving. *J. gen. Psychol.*, 1951, 44, 51-69.
- RHINE, R. J., The effect on problem solving of success or failure as a function of cue specificity. *J. exp. Psychol.*, 1957, 53, 121-125.
- RICHARDSON, J., & BERGUM, B. O., Distributed practice and rote learning in concept formation. *J. exp. Psychol.*, 1954, 47, 442-446.
- RILEY, D. A., Rote learning as a function of distribution of practice and the complexity of the situation. *J. exp. Psychol.*, 1952, 43, 88-95.
- RUSSELL, D. H., *Children's Thinking*. Boston: Ginn, 1956.
- SATO, T., An experimental study of problem solving in children and adults—solution of problems by principle learning. *Tohoku psychol. Fol.*, 1953, 13, 85-99.
- SAUGSTAD, P., Incidental memory and problem-solving. *Psychol. Rev.*, 1952, 59, 221-226.
- SAUGSTAD, P., Problem-solving as dependent upon availability of functions. *Brit. J. Psychol.*, 1955, 46, 191-198.
- SAUGSTAD, P., An analysis of Maier's pendulum problem. *J. exp. Psychol.*, 1957, 54, 168-179.
- SHAKLEE, A. B., & JONES, B. E., Distribution of practice prior to solution of a verbal reasoning problem. *J. exp. Psychol.*, 1953, 46, 429-434.
- SHAW, M. E., A comparison of individuals and small groups in the rational solution of complex problems. *Amer. J. Psychol.*, 1932, 44, 491-504.
- SOLLEY, C. M., Problem solving difficulty as a function of deviation of "meaning" of physical cues from expected "meaning." *J. gen. Psychol.*, 1957, 57, 165-171.
- STAATS, A. W., Verbal and instrumental response-hierarchies and their relationship to problem-solving. *Amer. J. Psychol.*, 1957, 70, 442-446.
- STOLUROW, L. M., BERGUM, B., HODGSON, T., & SILVA, J., The efficient course of action in "trouble shooting" as a joint function of probability and cost. *Educ. psychol. Measmt.*, 1955, 15, 462-477.

- STOLUROW, L. M., HODGSON, T. F., & SILVA, J., Transfer and retroaction effects of "association reversal" and "familiarization" training in trouble shooting. *Psychol. Monogr.*, 1956, 70 (12, Whole No. 419).
- SÜLLWOLD, F., Experimentelle Untersuchungen über die Rolle des Einfalles im Denkprozess. *Z. exp. angewand. Psychol.*, 1954, 2, 175-207.
- SWARTZ, P., Stimulus evolution in problem solving behavior: An inter-behavioral analysis. *Psychol. Rep.*, 1955, 1, 425-432.
- SZÉKELY, L., Zur Psychologie des inneren Verhaltens beim Lernen, Denken und Erfahren. *Theoria*, 1947, 13, 157-182.
- SZÉKELY, L., Knowledge and thinking. *Acta Psychol.*, 1950, 7, 1-24. (a)
- SZÉKELY, L., Productive processes in learning and thinking. *Acta Psychol.*, 1950, 7, 388-407. (b)
- TAYLOR, D. W., & FAUST, W. L., Twenty questions: Efficiency in problem solving as a function of size of group. *J. exp. Psychol.*, 1952, 44, 360-368.
- UNDERWOOD, B. J., *Experimental Psychology*. New York: Appleton-Century-Crofts, 1949.
- UNDERWOOD, B. J., An orientation for research on thinking. *Psychol. Rev.*, 1952, 59, 209-220.
- VAN DE GEER, J. P., *A Psychological Study of Problem Solving*. Haarlem: Uitgeverij De Toorts, 1957.
- VINACKE, W. E., *The Psychology of Thinking*. New York: McGraw-Hill, 1952.
- WEAVER, H. E., & MADDEN, E. H., "Direction" in problem solving. *J. Psychol.*, 1949, 27, 331-345.
- WERTHEIMER, M., *Productive Thinking*. New York: Harper, 1945.
- WIGGINS, J. G., Some relationships between stimulus structure and ambiguity in the solution of anagrams. *J. clin. Psychol.*, 1956, 12, 332-337.

JEROME S. BRUNER
Harvard University

The Act of Discovery*

Jerome S. Bruner, like B. F. Skinner, is a prominent experimental psychologist who has shown much interest in educational problems. The two men provide an interesting comparison

* Reprinted with the permission of the author and the publisher from the article of the same title, *Harvard Educational Review*, 31 (1961), 21-32.

because of their differing points of view; but these differences should not obscure what they share. To begin with, it may be quite significant that American education has turned away from philosophy, such as that of John Dewey, and toward a psychology based on objective research. Both Bruner and Skinner have attempted to translate psychological principles derived from laboratory research into research on classroom learning. Both men are now published widely in the scholarly journals as well as in the better popular magazines. Moreover, both men have expressed dissatisfaction with the drudgery and ineffectiveness of much present educational practice, which often resorts to verbal punishment or settles for dull resignation. In addition, both are concerned, as was Dewey, with allowing the learner a more active role in his learning: Skinner with his insistence on the importance of active recall and the "construction" of correct answers, and Bruner with his interest in active search and discovery. Both have studied the problem of improving the teaching of the traditional curriculum, so that subject matter has been restored to an honored position, although taught in highly unorthodox ways. Both are concerned with the problems of long-term retention and maximum transfer—what we have called efficient learning.

Their differences, however, may prove to be more significant than their similarities. Skinner seems to draw much more heavily upon laboratory research on animals than does Bruner, whose interest in perception, cognition, and social psychology has given his research a more humanistic flavor. Thinking, for Skinner, is a highly abstract concept which must be reduced to its specific component behaviors before we can teach it. For Bruner, however, thinking is a complex behavior, and it must be studied in all its complexity. To attempt to reduce it to bits of behavior would be to study something other than thinking. Skinner is not really concerned with *why* we learn, or in a theory of learning as such, but with *how* we learn and the practical conditions which promote learning. If some reward works in promoting learning, he would use it without conducting an inquiry into why it works. Bruner, however, is interested in such broad concepts as "hypothetical mode," "cumulative con-

structionism," and "intrinsic motivation"—terms which embrace great areas of behavior.

For the purposes of this chapter perhaps the most significant difference between the two psychologists lies in their conception of the learning act. The student is referred to Skinner's article on teaching machines (pp. 164-182) in which he indicates that learning should be a directed activity, so minutely and immediately directed that it requires the unfailing guidance of a machine. Learning materials must be reduced to steps so easy for the student to take that he rarely makes a mistake. It would almost seem that all the thinking has been done for the student beforehand by the programmer. Bruner, of course, is attempting to gather evidence to show that each learner, in one sense, must be his own programmer. In his active involvement in what he learns, he organizes information into cognitive structures which are meaningful and useful to *him*. To deprive the learner of the opportunity to do this is to rob him of the chance of achieving intellectual competence. The "shaping" of behavior which Skinner advocates is for Bruner only a means of acquiring learning at a low, concrete level. Bruner maintains that man's full intellectual capacities are developed by much more complex strategies of thinking.

The student should thoughtfully contrast Bruner's position in the following article with that of Skinner, presented earlier.

Maimonides, in his *Guide for the Perplexed*,¹ speaks of four forms of perfection that men might seek. The first and lowest form is perfection in the acquisition of worldly goods. The great philosopher dismisses such perfection on the ground that the possessions one acquires bear no meaningful relation to the possessor: "A great king may one morning find that there is no difference between him and the lowest person." A second perfection is of the body, its conformation and skills. Its failing is that it does not reflect on what is uniquely human about man: "he could [in any case] not be as

¹ Maimonides, *Guide for the Perplexed* (New York: Dover Publications, 1956).

strong as a mule." Moral perfection is the third, "the highest degree of excellency in man's character." Of this perfection Maimonides says: "Imagine a person being alone, and having no connection whatever with any other person; all his good moral principles are at rest, they are not required and give man no perfection whatever. These principles are only necessary and useful when man comes in contact with others." "The fourth kind of perfection is the true perfection of man; the possession of the highest intellectual faculties. . . ." In justification of his assertion, this extraordinary Spanish-Judaic philosopher urges: "Examine the first three kinds of perfection; you will find that if you possess them, they are not your property, but the property of others. . . . But the last kind of perfection is exclusively yours; no one else owns any part of it."

It is a conjecture much like that of Maimonides that leads me to examine the act of discovery in man's intellectual life. For if man's intellectual excellence is the most his own among his perfections, it is also the case that the most uniquely personal of all that he knows is that which he has discovered for himself. What difference does it make, then, that we encourage discovery in the learning of the young? Does it, as Maimonides would say, create a special and unique relation between knowledge possessed and the possessor? And what may such a unique relation do for a man—or for a child, if you will, for our concern is with the education of the young?

The immediate occasion for my concern with discovery—and I do not restrict discovery to the act of finding out something that before was unknown to mankind, but rather include all forms of obtaining knowledge for oneself by the use of one's own mind—the immediate occasion is the work of the various new curriculum projects that have grown up in America during the last six or seven years. For whether one speaks to mathematicians or physicists or historians, one encounters repeatedly an expression of faith in the powerful effects that come from permitting the student to put things together for himself, to be his own discoverer.

First, let it be clear what the act of discovery entails. It is rarely, on the frontier of knowledge or elsewhere, that new facts are "discovered" in the sense of being encountered as Newton suggested in the form of islands of truth in an uncharted sea of ignorance. Or if they appear to be discovered in this way, it is almost always thanks

to some happy hypotheses about where to navigate. Discovery, like surprise, favors the well prepared mind. In playing bridge, one is surprised by a hand with no honors in it at all and also by hands that are all in one suit. Yet all hands in bridge are equiprobable: one must know to be surprised. So too in discovery. The history of science is studded with examples of men "finding out" something and not knowing it. I shall operate on the assumption that discovery, whether by a schoolboy going it on his own or by a scientist cultivating the growing edge of his field, is in its essence a matter of rearranging or transforming evidence in such a way that one is enabled to go beyond the evidence so reassembled to additional new insights. It may well be that an additional fact or shred of evidence makes this larger transformation of evidence possible. But it is often not even dependent on new information.

It goes without saying that, left to himself, the child will go about discovering things for himself within limits. It also goes without saying that there are certain forms of child rearing, certain home atmospheres that lead some children to be their own discoverers more than other children. These are both topics of great interest, but I shall not be discussing them. Rather, I should like to confine myself to the consideration of discovery and "finding-out-for-oneself" within an educational setting—specifically the school. Our aim as teachers is to give our student as firm a grasp of a subject as we can, and to make him as autonomous and self-propelled a thinker as we can—one who will go along on his own after formal schooling has ended. I shall return in the end to the question of the kind of classroom and the style of teaching that encourages an attitude of wanting to discover. For purposes of orienting the discussion, however, I would like to make an overly simplified distinction between teaching that takes place in the *expository mode* and teaching that utilizes the *hypothetical mode*. In the former, the decisions concerning the mode and pace and style of exposition are principally determined by the teacher as expositor; the student is the listener. If I can put the matter in terms of structural linguistics, the speaker has a quite different set of decisions to make than the listener: the former has a wide choice of alternatives for structuring, he is anticipating paragraph content while the listener is still intent on the words, he is manipulating the content of the material by various

transformations, while the listener is quite unaware of these internal manipulations. In the hypothetical mode, the teacher and the student are in a more cooperative position with respect to what in linguistics would be called "speaker's decisions." The student is not a bench-bound listener, but is taking a part in the formulation and at times may play the principal role in it. He will be aware of alternatives and may even have an "as if" attitude toward these and, as he receives information he may evaluate it as it comes. One cannot describe the process in either mode with great precision as to detail, but I think the foregoing may serve to illustrate what is meant.

Consider now what benefit might be derived from the experience of learning through discoveries that one makes for oneself. I should like to discuss these under four headings: (1) The increase in intellectual potency, (2) the shift from extrinsic to intrinsic rewards, (3) learning the heuristics of discovering, and (4) the aid to memory processing.

1. *Intellectual potency.* If you will permit me, I would like to consider the difference between subjects in a highly constrained psychological experiment involving a two-choice apparatus. In order to win chips, they must depress a key either on the right or the left side of the machine. A pattern of payoff is designed such that, say, they will be paid off on the right side 70 per cent of the time, on the left 30 per cent, although this detail is not important. What is important is that the payoff sequence is arranged at random, and there is no pattern. I should like to contrast the behavior of subjects who think that there is some pattern to be found in the sequence—who think that regularities are discoverable—in contrast to subjects who think that things are happening quite by *chance*. The former group adopts what is called an "event-matching" strategy in which the number of responses given to each side is roughly equal to the proportion of times it pays off: in the present case R70 : L30. The group that believes there is no pattern very soon reverts to a much more primitive strategy wherein *all* responses are allocated to the side that has the greater payoff. A little arithmetic will show you that the lazy all-and-none strategy pays off more if indeed the environment is random: namely, they win seventy per cent of the time. The event-matching subjects win about 70% on the 70% payoff side

(or 49% of the time there) and 30% of the time on the side that pays off 30% of the time (another 9% for a total take-home wage of 58% in return for their labors of decision). But the world is not always or not even frequently random, and if one analyzes carefully what the event-matchers are doing, it turns out that they are trying out hypotheses one after the other, all of them containing a term such that they distribute bets on the two sides with a frequency to match the actual occurrence of events. If it should turn out that there is a pattern to be discovered, their payoff would become 100%. The other group would go on at the middling rate of 70%.

What has this to do with the subject at hand? For the person to search out and find regularities and relationships in his environment, he must be armed with an expectancy that there will be something to find and, once aroused by expectancy, he must devise ways of searching and finding. One of the chief enemies of such expectancy is the assumption that there is nothing one can find in the environment by way of regularity or relationship. In the experiment just cited, subjects often fall into a habitual attitude that there is either nothing to be found or that they can find a pattern by looking. There is an important sequel in behavior to the two attitudes, and to this I should like to turn now.

We have been conducting a series of experimental studies on a group of some seventy school children over the last four years. The studies have led us to distinguish an interesting dimension of cognitive activity that can be described as ranging from *episodic empiricism* at one end to *cumulative constructionism* at the other. The two attitudes in the choice experiments just cited are illustrative of the extremes of the dimension. I might mention some other illustrations. One of the experiments employs the game of Twenty Questions. A child—in this case he is between 10 and 12—is told that a car has gone off the road and hit a tree. He is to ask questions that can be answered by "yes" or "no" to discover the cause of the accident. After completing the problem, the same task is given him again, though he is told that the accident had a different cause this time. In all, the procedure is repeated four times. Children enjoy playing the game. They also differ quite markedly in the approach or strategy they bring to the task. There are various elements in the strategies employed. In the first place, one may distinguish clearly

between two types of questions asked: the one is designed for locating constraints in the problem, constraints that will eventually give shape to an hypothesis; the other is the hypothesis as question. It is the difference between, "Was there anything wrong with the driver?" and "Was the driver rushing to the doctor's office for an appointment and the car got out of control?" There are children who precede hypotheses with efforts to locate constraint and there are those who, to use our local slang, are "pot-shotters," who string out hypotheses non-cumulatively one after the other. A second element of strategy is its connectivity of information gathering: the extent to which questions asked utilize or ignore or violate information previously obtained. The questions asked by children tend to be organized in cycles, each cycle of questions usually being given over to the pursuit of some particular notion. Both within cycles and between cycles one can discern a marked difference on the connectivity of the child's performance. Needless to say, children who employ constraint location as a technique preliminary to the formulation of hypotheses tend to be far more connected in their harvesting of information. Persistence is another feature of strategy, a characteristic compounded of what appear to be two components: a sheer doggedness component, and a persistence that stems from the sequential organization that a child brings to the task. Doggedness is probably just animal spirits or the need for achievement—what has come to be called *n-ach*. Organized persistence is a maneuver for protecting our fragile cognitive apparatus from overload. The child who has flooded himself with disorganized information from unconnected hypotheses will become discouraged and confused sooner than the child who has shown a certain cunning in his strategy of getting information—a cunning whose principal component is the recognition that the value of information is not simply in getting it but in being able to carry it. The persistence of the organized child stems from his knowledge of how to organize questions in cycles, how to summarize things to himself, and the like.

Episodic empiricism is illustrated by information gathering that is unbound by prior constraints, that lacks connectivity, and that is deficient in organizational persistence. The opposite extreme is illustrated by an approach that is characterized by constraint sensi-

tivity, by connective maneuvers, and by organized persistence. Brute persistence seems to be one of those gifts from the gods that make people more exaggeratedly what they are.²

Before returning to the issue of discovery and its role in the development of thinking, let me say a word more about the ways in which information may get transformed when the problem solver has actively processed it. There is first of all a pragmatic question: what does it take to get information processed into a form best designed to fit some future use? Take an experiment by Zajonc³ as a case in point. He gives groups of subjects information of a controlled kind, some groups being told that their task is to transmit the information to others, others that it is merely to be kept in mind. In general, he finds more differentiation and organization of the information received with the intention of being transmitted than there is for information received passively. An active set leads to a transformation related to a task to be performed. The risk, to be sure, is in possible overspecialization of information processing that may lead to such a high degree of specific organization that information is lost for general use.

I would urge now in the spirit of an hypothesis that emphasis upon discovery in learning has precisely the effect upon the learner of leading him to be a constructionist, to organize what he is encountering in a manner not only designed to discover regularity and relatedness, but also to avoid the kind of information drift that fails to keep account of the uses to which information might have to be put. It is, if you will, a necessary condition for learning the variety of techniques of problem solving, of transforming information for better use, indeed for learning how to go about the very task of learning. Practice in discovering for oneself teaches one to acquire information in a way that makes that information more readily viable in problem solving. So goes the hypothesis. It is still in need of testing. But it is an hypothesis of such important human impli-

² I should also remark in passing that the two extremes also characterize concept attainment strategies as reported in *A Study of Thinking* by J. S. Bruner *et al.* (New York: J. Wiley, 1956). Successive scanning illustrates well what is meant here by episodic empiricism; conservative focussing is an example of cumulative constructionism.

³ R. B. Zajonc (Personal communication, 1957).

cations that we cannot afford not to test it—and testing will have to be in the schools.

2. *Intrinsic and extrinsic motives.* Much of the problem in leading a child to effective cognitive activity is to free him from the immediate control of environmental rewards and punishments. That is to say, learning that starts in response to the rewards of parental or teacher approval or the avoidance of failure can too readily develop a pattern in which the child is seeking cues as to how to conform to what is expected of him. We know from studies of children who tend to be early over-achievers in school that they are likely to be seekers after the "right way to do it" and that their capacity for transforming their learning into viable thought structures tends to be lower than children merely achieving at levels predicted by intelligence tests. Our tests on such children show them to be lower in analytic ability than those who are not conspicuous in overachievement. As we shall see later, they develop rote abilities and depend upon being able to "give back" what is expected rather than to make it into something that relates to the rest of their cognitive life. As Maimonides would say, their learning is not their own.

The hypothesis that I would propose here is that to the degree that one is able to approach learning as a task of discovering something rather than "learning about" it, to that degree will there be a tendency for the child to carry out his learning activities with the autonomy of self-reward or, more properly by reward that is discovery itself.

To those of you familiar with the battles of the last half-century in the field of motivation, the above hypothesis will be recognized as controversial. For the classic view of motivation in learning has been, until very recently, couched in terms of a theory of drives and reinforcement: that learning occurred by virtue of the fact that a response produced by a stimulus was followed by the reduction in a primary drive state. The doctrine is greatly extended by the idea of secondary reinforcement: any state associated even remotely with the reduction of a primary drive could also have the effect of producing learning. There has recently appeared a most searching and important criticism of this position, written by Professor Robert

White,⁴ reviewing the evidence of recently published animal studies, of work in the field of psychoanalysis, and of research on the development of cognitive processes in children. Professor White comes to the conclusion, quite rightly I think, that the drive-reduction model of learning runs counter to too many important phenomena of learning and development to be either regarded as general in its applicability or even correct in its general approach. Let me summarize some of his principal conclusions and explore their applicability to the hypothesis stated above.

I now propose that we gather the various kinds of behavior just mentioned, all of which have to do with effective interaction with the environment, under the general heading of competence. According to Webster, competence means fitness or ability, and the suggested synonyms include capability, capacity, efficiency, proficiency, and skill. It is therefore a suitable word to describe such things as grasping and exploring, crawling and walking, attention and perception, language and thinking, manipulating and changing the surroundings, all of which promote an effective—a competent—interaction with the environment. It is true of course, that maturation plays a part in all these developments, but this part is heavily overshadowed by learning in all the more complex accomplishments like speech or skilled manipulation. I shall argue that it is necessary to make competence a motivational concept; there is *competence motivation* as well as competence in its more familiar sense of achieved capacity. The behavior that leads to the building up of effective grasping, handling, and letting go of objects, to take one example, is not random behavior that is produced by an overflow of energy. It is directed, selective, and persistent, and it continues not because it serves primary drives, which indeed it cannot serve until it is almost perfected, but because it satisfies an intrinsic need to deal with the environment.⁵

I am suggesting that there are forms of activity that serve to enlist and develop the competence motive, that serve to make it the driving force behind behavior. I should like to add to White's general premise that the *exercise* of competence motives has the effect of strengthening the degree to which they gain control over behavior and thereby reduce the effects of extrinsic rewards or drive gratification.

⁴ R. W. White, "Motivation Reconsidered: The Concept of Competence," *Psychological Review*, LXVI (1959), 297-333.

⁵ *Ibid.*, pp. 317-18.

The brilliant Russian psychologist Vigotsky⁶ characterizes the growth of thought processes as starting with a dialogue of speech and gesture between child and parent; autonomous thinking begins at the stage when the child is first able to internalize these conversations and "run them off" himself. This is a typical sequence in the development of competence. So too in instruction. The narrative of teaching is of the order of the conversation. The next move in the development of competence is the internalization of the narrative and its "rules of generation" so that the child is now capable of running off the narrative on his own. The hypothetical mode in teaching by encouraging the child to participate in "speaker's decisions" speeds this process along. Once internalization has occurred, the child is in a vastly improved position from several obvious points of view—notably that he is able to go beyond the information he has been given to generate additional ideas that can either be checked immediately from experience or can, at least, be used as a basis for formulating reasonable hypotheses. But over and beyond that, the child is now in a position to experience success and failure not as reward and punishment, but as information. For when the task is his own rather than a matter of matching environmental demands, he becomes his own paymaster in a certain measure. Seeking to gain control over his environment, he can now treat success as indicating that he is on the right track, failure as indicating he is on the wrong one.

In the end, this development has the effect of freeing learning from immediate stimulus control. When learning in the short run leads only to pellets of this or that rather than to mastery in the long run, then behavior can be readily "shaped" by extrinsic rewards. When behavior becomes more long-range and competence-oriented, it comes under the control of more complex cognitive structures, plans and the like, and operates more from the inside out. It is interesting that even Pavlov, whose early account of the learning process was based entirely on a notion of stimulus control of behavior through the conditioning mechanism in which, through contiguity a new conditioned stimulus was substituted for an old unconditioned stimulus by the mechanism of stimulus substitution,

⁶ L. S. Vigotsky, *Thinking and Speech* (Moscow, 1934).

that even Pavlov recognized his account as insufficient to deal with higher forms of learning. To supplement the account, he introduced the idea of the "second signalling system," with central importance placed on symbolic systems such as language in mediating and giving shape to mental life. Or as Luria⁷ has put it, "the first signal system [is] concerned with directly perceived stimuli, the second with systems of verbal elaboration." Luria, commenting on the importance of the transition from first to second signal system, says: "It would be mistaken to suppose that verbal intercourse with adults merely changes the contents of the child's conscious activity without changing its form. . . . The word has a basic function not only because it indicates a corresponding object in the external world, but also because it abstracts, isolates the necessary signal, generalizes perceived signals and relates them to certain categories; it is this systematization of direct experience that makes the role of the word in the formation of mental processes so exceptionally important."⁸

It is interesting that the final rejection of the universality of the doctrine of reinforcement in direct conditioning came from some of Pavlov's own students. Ivanov-Smolensky⁹ and Krasnogorsky¹⁰ published papers showing the manner in which symbolized linguistic messages could take over the place of the unconditioned stimulus and of the unconditioned response (gratification of hunger) in children. In all instances, they speak of these as *replacements* of lower, first-system mental or neural processes by higher order or second-system controls. A strange irony, then, that Russian psychology that gave us the notion of the conditioned response and the assumption that higher order activities are built up out of colligations or structurings of such primitive units, rejected this notion while much of American learning psychology has stayed until quite recently within the early Pavlovian fold (see, for example, a recent

⁷ A. L. Luria, "The Directive Function of Speech in Development and Dissolution," *Word*, XV (1959), 341-464.

⁸ *Ibid.*, p. 12.

⁹ A. G. Ivanov-Smolensky, "Concerning the Study of the Joint Activity of the First and Second Signal Systems," *Journal of Higher Nervous Activity*, I (1951), 1.

¹⁰ N. D. Krasnogorsky, *Studies of Higher Nervous Activity in Animals and in Man*, Vol. I (Moscow, 1954).

article by Spence¹¹ in the *Harvard Educational Review* or Skinner's treatment of language¹² and the attacks that have been made upon it by linguists such as Chomsky¹³ who have become concerned with the relation of language and cognitive activity). What is the more interesting is that Russian pedagogical theory has become deeply influenced by this new trend and is now placing much stress upon the importance of building up a more active symbolical approach to problem solving among children.

To sum up the matter of the control of learning, then, I am proposing that the degree to which competence or mastery motives come to control behavior, to that degree the role of reinforcement or "extrinsic pleasure" wanes in shaping behavior. The child comes to manipulate his environment more actively and achieves his gratification from coping with problems. Symbolic modes of representing and transforming the environment arise and the importance of stimulus-response-reward sequences declines. To use the metaphor that David Riesman developed in a quite different context, mental life moves from a state of outer-directedness in which the fortuity of stimuli and reinforcement are crucial to a state of inner-directedness in which the growth and maintenance of mastery become central and dominant.

3. *Learning the heuristics of discovery.* Lincoln Steffens,¹⁴ reflecting in his *Autobiography* on his under graduate education at Berkeley, comments that his schooling was overly specialized on learning about the known and that too little attention was given to the task of finding out about what was not known. But how does one train a student in the techniques of discovery? Again I would like to offer some hypotheses. There are many ways of coming to the arts of inquiry. One of them is by careful study of its formalization in logic, statistics, mathematics, and the like. If a person is going to pursue inquiry as a way of life, particularly in the sciences, certainly such study is essential. Yet, whoever has taught kindergarten

¹¹ K. W. Spence, "The Relation of Learning Theory to the Technique of Education," *Harvard Educational Review*, XXIX (1959), 84-95.

¹² B. F. Skinner, *Verbal Behavior* (New York: Appleton-Century-Crofts, 1957).

¹³ N. Chomsky, *Syntactic Structure* (The Hague, The Netherlands: Mouton & Co., 1957).

¹⁴ L. Steffens, *Autobiography of Lincoln Steffens* (New York: Harcourt, Brace & World, 1931).

and the early primary grades or has had graduate students working with him on their theses—I choose the two extremes for they are both periods of intense inquiry—knows that an understanding of the formal aspect of inquiry is not sufficient. There appear to be, rather, a series of activities and attitudes, some directly related to a particular subject and some of them fairly generalized, that go with inquiry and research. These have to do with the *process* of trying to find out something and while they provide no guarantee that the *product* will be any *great* discovery, their absence is likely to lead to awkwardness or aridity or confusion. How difficult it is to describe these matters—the heuristics of inquiry. There is one set of attitudes or ways of doing that has to do with sensing the relevance of variables—how to avoid getting stuck with edge effects and getting instead to the big sources of variance. Partly this gift comes from intuitive familiarity with a range of phenomena, sheer “knowing the stuff.” But it also comes out of a sense of what things among an ensemble of things “smell right” in the sense of being of the right order of magnitude or scope or severity.

The English philosopher Weldon describes problem solving in an interesting and picturesque way. He distinguishes between difficulties, puzzles, and problems. We solve a problem or make a discovery when we impose a puzzle form on to a difficulty that converts it into a problem that can be solved in such a way that it gets us where we want to be. That is to say, we recast the difficulty into a form that we know how to work with, then work it. Much of what we speak of as discovery consists of knowing how to impose what kind of form on various kinds of difficulties. A small part but a crucial part of discovery of the highest order is to invent and develop models or “puzzle forms” that can be imposed on difficulties with good effect. It is in this area that the truly powerful mind shines. But it is interesting to what degree perfectly ordinary people can, given the benefit of instruction, construct quite interesting and what, a century ago, would have been considered greatly original models.

Now to the hypothesis. It is my hunch that it is only through the exercise of problem solving and the effort of discovery that one learns the working heuristic of discovery, and the more one has practice, the more likely is one to generalize what one has learned

into a style of problem solving or inquiry that serves for any kind of task one may encounter—or almost any kind of task. I think the matter is self-evident, but what is unclear is what kinds of training and teaching produce the best effects. How do we teach a child to, say, cut his losses but at the same time be persistent in trying out an idea; to risk forming an early hunch without at the same time formulating one so early and with so little evidence as to be stuck with it waiting for appropriate evidence to materialize; to pose good testable guesses that are neither too brittle nor too sinuously incorrigible; etc., etc. Practice in inquiry, in trying to figure out things for oneself is indeed what is needed, but in what form? Of only one thing I am convinced. I have never seen anybody improve in the art and technique of inquiry by any means other than engaging in inquiry.

4. *Conservation of memory.* I should like to take what some psychologists might consider a rather drastic view of the memory process. It is a view that in large measure derives from the work of my colleague, Professor George Miller.¹⁵ Its first premise is that the principal problem of human memory is not storage, but retrieval. In spite of the biological unlikeliness of it, we seem to be able to store a huge quantity of information—perhaps not a full tape recording, though at times it seems we even do that, but a great sufficiency of impressions. We may infer this from the fact that recognition (i.e., recall with the aid of maximum prompts) is so extraordinarily good in human beings—particularly in comparison with spontaneous recall where, so to speak, we must get out stored information without external aids or prompts. The key to retrieval is organization or, in even simpler terms, knowing where to find information and how to get there.

Let me illustrate the point with a simple experiment. We present pairs of words to twelve-year-old children. One group is simply told to remember the pairs, that they will be asked to repeat them later. Another is told to remember them by producing a word or idea that will tie the pair together in a way that will make sense to them. A third group is given the mediators used by the second group

¹⁵ G. A. Miller, "The Magical Number Seven, Plus or Minus Two," *Psychological Review*, LXIII (1956), 81-97.

when presented with the pairs to aid them in tying the pairs into working units. The word pairs include such juxtapositions as "chair-forest," "sidewalk-square," and the like. One can distinguish three styles of mediators and children can be scaled in terms of their relative preference for each: *generic mediation* in which a pair is tied together by a superordinate idea: "chair and forest are both made of wood"; *thematic mediation* in which the two terms are imbedded in a theme or little story: "the lost child sat on a chair in the middle of the forest"; and *part-whole mediation* where "chairs are made from trees in the forest" is typical. Now, the chief result, as you would all predict, is that children who provide their own mediators do best—indeed, one time through a set of thirty pairs, they recover up to 95% of the second words when presented with the first ones of the pairs, whereas the uninstructed children reach a maximum of less than 50% recovered. Interestingly enough, children do best in recovering materials tied together by the form of mediator they most often use.

One can cite a myriad of findings to indicate that any organization of information that reduces the aggregate complexity of material by imbedding it into a cognitive structure a person has constructed will make that material more accessible for retrieval. In short, we may say that the process of memory, looked at from the retrieval side, is also a process of problem solving: how can material be "placed" in memory so that it can be got on demand?

We can take as a point of departure the example of the children who developed their own technique for relating the members of each word pair. You will recall that they did better than the children who were given by exposition the mediators they had developed. Let me suggest that in general, material that is organized in terms of a person's own interests and cognitive structures is material that has the best chance of being accessible in memory. That is to say, it is more likely to be placed along routes that are connected to one's own ways of intellectual travel.

In sum, the very attitudes and activities that characterize "figuring out" or "discovering" things for oneself also seems to have the effect of making material more readily accessible in memory.

HOWARD H. KENDLER, *New York University*

TRACY S. KENDLER, *Barnard College*

Inferential Behavior in Preschool Children *

When does the child reach the age of reason? In Christian theology the child does not bear full moral responsibility until he is seven. The law remains rather skeptical about his rational powers until he reaches twenty-one. Despite these efforts to fix a determinant age, the work of developmental psychologists indicates that reason is not a suddenly dawning event; the child becomes progressively more reasonable as he grows older. The investigation by the Kendlers reported here indicates that very young children are capable of fairly sophisticated thinking.

In reviewing the following article the student may consider these questions: (1) What particular inference were these children asked to make? (2) The Kendlers claim that inferential behavior explains the superiority of the experimental groups over the others. What physical-spatial cue might the children have been using (see Duncan, p. 218)? (3) The authors take issue with Maier. How does this disagreement reflect what Duncan has described as the present state of research on human problem solving? (4) What implications does this article have for curricular reform in the elementary schools? What new learning experiences would you introduce into the early grades?

This study was designed primarily to determine whether children, three to four years old, would be capable of inferential behavior. The experimental situation used was similar to the one suggested

* Reprinted with the permission of the author and the American Psychological Association from the article of the same title, *Journal of Experimental Psychology*, 51 (1956), 311-314. Footnotes are omitted.

by Hull to serve as a test of his theoretical account of the reasoning behavior of rats as reported by Maier.

In our experimental procedure, which is diagrammatically represented in Fig. 1, *S* received three separate experiences. The *S* learned to pull A to get subgoal B, to pull X to get subgoal Y, and to pull B to get major goal G. In the test trial *S* was presented with A and X and told to choose the one that would obtain G. If he chose A, his behavior was designated as inferential.

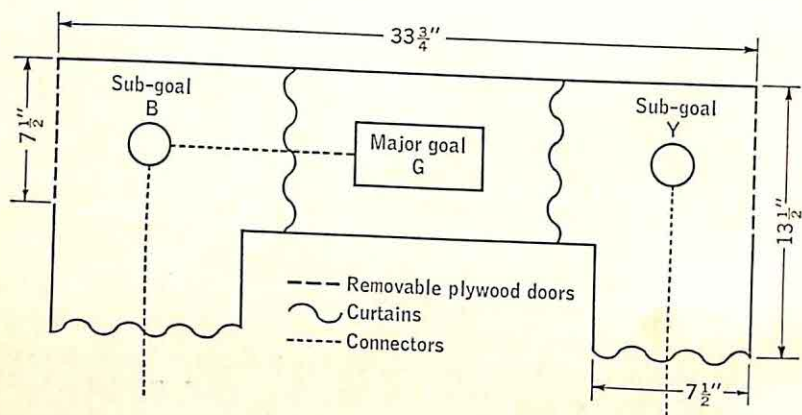


Fig. 1. Floor Plan of Apparatus.

In addition to discovering whether inferential behavior would occur, we were also interested in discovering whether the order of presentation of the three distinct experiences would influence the amount of inferential behavior. Hull's analysis, which presupposed the use of inarticulate *Ss*, led to the prediction that order would influence the amount of inference. Reasoning that inferential behavior is mediated by the mechanism of the anticipatory goal response, Hull analyzed the process in the following manner: An r_p is developed to the major goal G which moves back to B and from there back to A. When presented with a choice between A and X, *S* should choose A because it evokes anticipatory goal responses for the major goal, which in turn produce cues which have become conditioned to responses in the direction of B and G. According to this analysis there should be more inference when the B-G experience

precedes A-B than when it follows it, because in order for A to acquire the capacity to evoke an r_g for the major goal, B must have had an opportunity to become conditioned to the r_g of the major goal.

Although the present experiment was not primarily designed to test this hypothesis, and consequently does not provide a definitive test, it does throw some light on the effect of order of presentation of experiences on inferential behavior in preschool children.

METHOD

Experimental design. The design included four experimental and four control groups with 16 Ss in each group. Each experimental group had the three experiences in a different order. As shown in Table 1,

Table 1

DIAGRAMMATIC REPRESENTATION OF THE EXPERIMENTAL DESIGN

Groups	Order of Experiences			
	1	2	3	4
Experimental	A-B, X-Y, B-G	X-Y, A-B, B-G	B-G, A-B, X-Y	B-G, X-Y, A-B
Control	A-B, X-Y, G	X-Y, A-B, G	G, A-B, X-Y	G, X-Y, A-B

B-G precedes A-B in two orders and follows it in two others. The control groups differed from the experimental groups only in that they attained G without ever experiencing B-G. These groups were included to control for possible effects, other than the B-G experience, that might influence Ss choices during the test trials, e.g., the effect of which of the three experiences terminated the sequence.

Subjects. The Ss were 128 (68 M and 60 F) children between 34 and 59 mo. of age. They were drawn from three private nursery schools. Sixteen Ss were assigned at random to each experimental group, with the limiting condition that eight Ss in each group were drawn from children between 48 and 59 mo. old, and eight from children between 34 and 47 mo. old. The experimental groups were not equated for sex.

Apparatus. The apparatus (Fig. 1) was a portable maze-like structure made of 1/4-in. plywood in which the goals were pulled along the alleys by means of strings until they appeared at the appropriate openings. The entire apparatus was covered by a fitted plywood section so that S saw a gray box with two curtains at A and X. The covering plywood was jointed in a manner that permitted the top to be lifted to make the necessary adjustments of goal objects while screening E's

manipulation from S. At B and Y there were removable plywood "doors."

The minor goal objects at B and Y were either a small stuffed red ladybug or a small stuffed gray chicken. At G, the major goal, was a foreign toy sports car (Schuco-Akustico 2002) that ran when its brake was released. It also had a button in the center of its steering wheel which, when pressed, emitted a realistic car-horn sound.

Procedure. In each school a room was assigned to E for the experiment. The apparatus was always placed on the floor. The S sat on the floor in front of and facing A and X. The E sat on the opposite side of the apparatus facing S. The experimental task was introduced to S as a game.

Each experimental S was presented with each experience twice in succession but in different orders, as indicated in Table 1. For example, if S was in Exp. Group 1, E first put out the ribbon at A and told S to pull it. The S pulled and the stuffed ladybug emerged through the curtains at A. Although a string was clipped to both subgoals to prevent them from being pulled more than a few inches past the apparatus, there was sufficient play in it to allow S to handle the stuffed toy. After a few moments of handling, the ladybug was retracted by E. Then the ribbon was again put out past the curtains at A and the procedure described above was repeated. When the two A-B experiences were completed, the gold chain was set out past the curtains at X. The S pulled and out came the chicken, which was handled before it was retracted. This was repeated to provide two X-Y experiences.

For the B-G experience the ladybug was connected by a string to the car and the plywood door at B was removed. The S was instructed to come around to B and "peek in." He could see the ladybug but the car was hidden by the curtain between B and G. The S was then told to pull out the bug. When S pulled the bug, he also pulled out the car. He was permitted to play with the car for about one minute. During this time he made it go around the room once and tooted the horn two or three times. Then the ladybug and the car were replaced in the apparatus, hooked up again, and the procedure was repeated.

The next trial was the test trial. Both the ribbon at A and the chain at X were set out simultaneously. The S was instructed to pull *one* of them, the one that would get him the car.

Although the order of presentation of the experiences differed for the remainder of the experimental groups, each experience was conducted in precisely the manner described except that the side of the apparatus (right or left) on which A was placed, the subgoals (chicken or ladybug) and the character of the connectors at A and X (ribbon or chain) were counterbalanced for each experimental and control group.

The control groups differed from the experimental groups only in

that the B-G experience was absent. In order to keep both the experience of pulling and the experience with the major goal constant, the control groups were presented with a small gray box, $8\frac{1}{2}$ in. long $\times$ 3 in. high $\times$ 4 in. wide, in which the car was placed. Attached to the car was a green plastic string. The child pulled this and the car emerged from behind the curtains that covered the opening. He was permitted to play with the car in the same way that the experimental Ss were. The remainder of the procedure was identical with that of the experimental groups.

RESULTS AND DISCUSSION

The test-trial results for all experimental and control groups are presented in Table 2. Evidently the experimental procedure is

Table 2
NUMBER AND PERCENTAGE OF Ss WHO MADE INFERENTIAL CHOICE ON
TEST TRIAL FOR EACH EXPERIMENTAL AND CONTROL GROUP

Group	Order	Experimental Groups		Control Groups	
		N	%	N	%
1	A-B, X-Y, B-G (or G)	12	75.0	7	43.8
2	X-Y, A-B, B-G (or G)	14	87.5	7	43.8
3	B-G (or G), A-B, X-Y	13	81.2	12	75.0
4	B-G (or G), X-Y, A-B	7	43.8	4	25.0
Total		46	71.9	30	46.9

capable of producing inference in nursery school children. The difference in frequency between the total experimental and total control groups yielded a χ^2 of 8.29, $P = < .01$. There was no significant difference between the amount of inference demonstrated by the two age groups.

These results are in opposition to a conclusion reached by Maier based upon his study of reasoning in children which was reported in 1936. His results showed that the preschool children in his experiment did not perform significantly better than chance when required to combine past experiences to reach a goal. Maier concluded that "... the ability to combine the essentials of two isolated experiences in such a manner as to reach a goal is rather late in maturing. It is rarely developed to a marked extent in children

below six years of age." The experimental results on which Maier's conclusion was based are not necessarily inconsistent with the present results. The task Maier used was more complex than our relatively simple and direct procedure. The comparison between these two sets of data suggests the need to consider more carefully the nature of the experimental task before reaching such general conclusions.

Table 2 also shows that when either the A-B or X-Y experience immediately preceded the test trial, Ss tended to choose the opposite side on the test trial. This tendency occurred in both the experimental and control groups (3 and 4).

Table 3 presents comparisons designed to explore the effects of

Table 3

NUMBER AND PERCENTAGE OF INFERENTIAL RESPONSES ON TEST TRIAL
ACCORDING TO POSITION OF THE B-G EXPERIENCE

Position of B-G	Experimental Groups		Control Groups		χ^2	p
	N	%	N	%		
B-G (or G) last (groups 1 and 2 combined)	26	81.2	14	43.8	9.60	<.01
B-G (or G) first (groups 3 and 4 combined)	20	62.5	16	50.0	1.01	.50

the relative positions of B-G and A-B. In this analysis group results were combined to provide a sufficiently large N for a valid test of statistical reliability. These results indicate that a significant amount of inference does occur when A-B precedes B-G but not when B-G occurs initially. These results fail to support Hull's hypothesis that a sequence of experiences in which B-G precedes A-B should produce more inferential behavior than when the order is reversed. We feel, however, that a more stringent test entailing direct comparisons between experimental groups, varying the B-G and A-B orders while *holding constant the last experience*, is required before a clear-cut conclusion can be drawn about the relevance of Hull's hypothesis to articulate organisms.

SUMMARY

An experiment was conducted to determine whether children, 3-4 yr. of age, were capable of making an inferential response on the basis

of past training. The experimental Ss received three separate experiences in a maze-like situation. Two of these experiences (A-B and X-Y) led to subgoals while the third experience (B-G) led to a major goal. The subgoal in the A-G experience served as the start of the B-G experience. The control Ss had a similar set of three experiences, with the exception that the experience involving G was not preceded by B. When Ss were given a choice between A and X with instructions to choose the alternative which led to G, significantly more experimental Ss than control Ss chose A. These results were interpreted as demonstrating inferential behavior in preschool children. The data also suggested that inferential behavior was more likely to occur when the A-B experience preceded B-G than when B-G occurred prior to A-B.

References

- HULL, C. L., *A Behavior System*. New Haven: Yale Univ. Press, 1952.
- LEWIS, D., & BURKE, C. J., The use and misuse of the chi-square test. *Psychol. Bull.*, 1949, 46, 433-489.
- MAIER, N. R. F., Reasoning in white rats. *Comp. psychol. Monogr.*, 1929, No. 29.
- MAIER, N. R. F., Reasoning in children. *J. comp. Psychol.*, 1936, 21, 357-366.

BERT Y. KERSH

Teaching Research, Oregon State System of Higher Education

The Motivating Effect of Learning by Directed Discovery*

How much guidance should the teacher give to the student if he wants the student to "understand" the materials and ideas at hand and not merely to parrot them back? The question was raised in the introduction to this chapter and in the review by Duncan (pp. 216-219), and it is experimentally investigated in the article which follows.

* Published with the permission of the author.

The ultimate test of efficient learning in the classroom is always some future performance of the student. In psychological parlance this is the question of retention (remembering) and transfer, or the use of the learning under the appropriate conditions (pp. 31-33). Accordingly, Kersh has employed the criteria of retention and transfer in this study.

Kersh claims that his evidence (as well as that of other investigators) does not support all of the claims which Bruner makes in his article on discovery (pp. 256-270). (1) For what parts of his investigation does this seem true? (2) Is the superiority of the guided discovery group due to the training method? To what does Kersh attribute its superiority? (3) Kersh was surprised that the rote learning group did so very well. Why might rote learning succeed where discovery learning fails? (4) How might the superiority of the guided discovery group be explained by learning set? (5) For what educational objectives might the discovery method be most useful?

Advocates of the process of learning by directed discovery claim a number of advantages, most of which are included in a recent article by Bruner (1961). He has suggested that learning by discovery benefits the learner in four ways: it (a) increases the learner's ability to learn related material, (b) fosters an interest in the activity itself rather than in the rewards which may follow from the learning, (c) develops ability to approach problems in a way that will more likely lead to a solution, and (d) tends to make the material that is learned easier to retrieve or reconstruct.

Research evidence does not entirely support Bruner's arguments. One of the more recent reviewers, Ausubel (1961), concludes "that most of the reasonably well-controlled studies report negative findings." However, as is true in other areas of research, the evidence is somewhat equivocal, partly because it is difficult to equate studies in terms of the amount and kind of direction that is provided. The experimental subjects rarely if ever are required to learn completely without help, and the kind of help provided commonly differs. Consequently, there are studies which appear to be somewhat contradic-

tory, such as Craig's (1956), in which the "directed" group learned and retained significantly more principles than the "independent" group, and Kittell's (1957), in which the group which received an intermediate amount of guidance was superior in learning, retention, and transfer to groups receiving either more or less direction. It has been suggested that the "intermediate" amount of guidance provided by Kittell may have actually exceeded the amount Craig provided to his directed group (Ausubel, 1961).

One of the few studies that forced learners to discover almost entirely without help provides data in support of the discovery process (Kersh, 1958). The contrasting directed treatment groups were superior in learning rate and immediate recall, but the "no help" group was superior in terms of retention and transfer after a period of approximately one month following the learning period. No evidence was produced to indicate that the no help group understood the rules better. Instead, an explanation was offered in terms of practice. On the basis of a subjective analysis of the subject's comments written on the retests and reported to the experimenter, it was concluded that the learners were motivated to continue the learning process or to continue practicing the task after the formal learning period.

The present experiment was designed to provide formal data concerning the motivating power in question.

HYPOTHESIS

Each subject had the task of learning the following two rules of addition:

1. Odd Numbers rule. The sum of any series of consecutive odd numbers beginning with 1 is equal to the square of the number of figures in the series. (For example, 1, 3, 5, 7, is such a series; there are four numbers, so 4×4 is 16, the sum.)
2. Constant Difference rule. The sum of any series of numbers in which the difference between the numbers is constant is equal to one-half the product of the number of figures and the sum of the first and last numbers. (For example, 2, 3, 4, 5, is such a series; 2 and 5 are 7; there are four figures, so 4×7 is 28; half of 28 is 14 which is the sum.)

The rules can be learned by simple memorization of the task procedure as above. Further, the learner can become cognizant of certain relations which these rules bear to geometrical and arithmetical concepts, in which case it is assumed that his learning will be more meaningful. The definition of meaning, as well as the geometrical and arithmetical relationships referred to are identified in a previous publication (Kersh, 1958). In the hypothesis statement below, the term "relationships" refers specifically to those in the reference cited above and generally to comparable relationships in related tasks.

As will be explained below, the experimental treatments in the present study differed primarily with respect to the extent of the external direction provided the subjects in learning the relationships referred to above. The present experiment was designed to test the following hypothesis.

To the extent that the external direction provided to the learner is lessened during his attempts to discover the relationships which are considered essential to the understanding of a cognitive task: (a) the learner will tend to use the learned material more frequently after the learning period (i.e., to extend the practice period voluntarily) and, as a result, (b) he will remember it longer and transfer his learning more effectively.

It should be noted that the hypothesis is written in two parts and that the second is dependent upon the first.

PROCEDURE

A total of 90 high school geometry students was utilized, having been selected from a larger group on the basis of a pretest covering the arithmetical and geometrical concepts and procedures that were considered essential prerequisites to the tasks used in the experiment. The entire sample was then taught the two rules of addition given above by being simply told the rules and given practice in their application. They were taught by a programed booklet procedure to the same criterion, six successive applications of each of the two rules. Thereafter, the subjects were divided at random into three main groups of 30 each, and each group was treated differently.

One group, called the Directed Learning group, was taught the rules and their explanation entirely by a programed learning technique. Each subject learned from a booklet in which the learning process was broken down into smaller steps, and answers to questions or solutions

to problems were revealed to the subject whether he responded correctly or not.

A second group, called the Guided Discovery group, was required to discover the explanation with guidance from the experimenter. The subjects in the Guided Discovery group were taught tutorially using a form of Socratic questioning which required each subject to perform specific algebraic manipulations and to make inferences without help. The guidance was a practical expedient, since it was necessary to control between groups the quality and quantity of the relationships used in explaining the rules.

The final group was called, appropriately, the Rote Learning group since the explanation for the rules was omitted. This treatment was incorporated in the research design primarily as the control for "meaningful" learning.

After the learning period of the experiment, a test of recall and transfer was given to subgroups of each treatment group after 3 days, 2 weeks, and 6 weeks. For this purpose each of the three main groups was divided into three subgroups of 10 each.

The test consisted of two problems and a short questionnaire. The problems were given first with instructions to show all work including scratchwork. The two test problems were as follows:

1. John's employer agrees to pay him \$1.00 for his first day of work and increase his pay by \$2.00 each day. How much will he receive for the first month's work if he works all 30 days?

2. A man is left a sum of money by an eccentric relative. The will states that he will receive \$10.00 the first month and that each successive monthly payment will be increased by \$5.00 (i.e., he will receive \$10.00 the first month, \$15.00 the second month, \$20.00 the third month, etc.). His monthly payment at the end of four years is \$245.00. What is the total amount he has been paid by that time?

The questionnaire asked the subject to state each rule, using examples if necessary, and to report whether or not he made use of the rules after the formal learning period.

RESULTS

The number of subjects in each group who used the appropriate rule in an acceptable way on the test was employed as the index of transfer. Acceptable use of a rule for the first test problem meant the use of either rule to obtain the solution; for the second test problem, only the Constant Difference rule was acceptable. Computational accuracy was not required.

The number of subjects in each group who wrote an acceptable

statement of each rule was used as a measure of pure retention. To be acceptable, each subject's statement had to be complete and accurate, but not necessarily in the same words as the original statements. Errors in spelling or grammar were overlooked.

Table 1 presents the number who used and stated the rules in the acceptable way on the test problems. A total of 90 subjects served as the basis for the data in Table 1, 10 subjects per cell.

In the statistical analysis, use was made of a chi square technique devised by Li (1957). The data included under each of the columns of Table 1 were envisioned as a 2×9 contingency table, with 8 *df*. Four separate analyses were then conducted, each of which broke down the chi square into the following components: (a) differences between teaching treatments (2 *df*), (b) differences between test periods (2 *df*), and (c) differences attributed to interaction of treatments and time period (4 *df*).

Table 1
NUMBER OF SUBJECTS (OF 10 IN EACH CELL) WHO USED AND
STATED THE RULES CORRECTLY ON THE RETEST

Treatment Groups	Used Rules		Stated Rules	
	<i>Odd Numbers^a</i>	<i>Constant Difference</i>	<i>Odd Numbers^a</i>	<i>Constant Difference^a</i>
	1	2	3	4
Rote Learning				
3 days	7	7	7	9
2 weeks	7	6	2	6
6 weeks	4	4	0	3
Guided Discovery:				
3 days	6	6	8	9
2 weeks	3	5	3	4
6 weeks	2	3	3	3
Directed Learning:				
3 days	4	3	3	4
2 weeks	4	3	1	3
6 weeks	0	3	1	1

^a Differences between treatment groups and between test periods significant by chi square at or beyond .05 level.

None of the interaction effects was significant, indicating that the rate of forgetting did not differ significantly across the teaching

treatment groups. A trend analysis of the test data indicated also that the rate of forgetting was constant for all groups (Li, 1957, pp. 226-233).

Otherwise, as pointed out by the footnote references in Table 1, the differences between treatment groups and between test periods were found to be significant for all columns except that headed "Constant Difference 2," for which the observed differences were found not to be reliable.

Perhaps the most striking finding in the present study is that the Rote Learning group was found to be consistently superior in every respect to the other treatment groups. Although this completely unanticipated finding has no direct bearing on the hypothesis in question, it does nevertheless bear clearly upon the related question of "meaningful vs. mechanical" learning. This finding will be discussed in a subsequent section.

Strictly speaking, the hypothesis which the present experiment was designed to test involves only the Guided Discovery and Directed Learning treatments. To support the major hypothesis, the data should have shown that the subjects comprising the Guided Discovery group used the rules after the learning period more frequently than the subjects in the Directed Learning group; and, if so, that the former remembered and transferred the rules more effectively than the latter.

With respect to the frequency of using the rules after the learning period, the results do support the hypothesis. Although the number of subjects in each group who reported that they did use the rules was very small, the difference between the frequency patterns of the two groups in question is statistically significant. Eleven subjects of the 30 in the Guided Discovery group reported that they had used the rules as compared with two subjects in the Directed Learning group. In the Rote Learning group, six subjects of 30 reported in the affirmative.

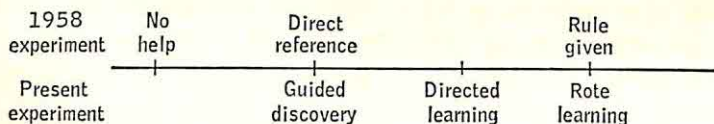
With respect to the relative permanence of the retention and increased transfer effects, the results also support the hypothesis. The Guided Discovery group is clearly superior to the Directed Learning group 3 days after the learning period, and since the rate of forgetting may be presumed to be approximately the same for each treatment group (see statistical analysis above), their initial superiority remains after 6 weeks.

DISCUSSION

The data from this present experiment do not support the generalization that learning by a process which involves discovery is necessarily superior to learning by more highly directed processes. Indeed, these data suggest that under certain conditions of learning, highly formalized "lecture-drill" techniques, ordinarily considered sterile and meaningless, produce better results than techniques which attempt to develop "understanding."

One explanation for the present results is that they reflect a simple and well known phenomenon, retroactive inhibition. The experimental efforts to inject meaning into the rules amounted to following their initial rote learning with a closely related and complex learning task; thus the Rote Learning group may have surpassed other groups simply because retention among the latter was inhibited by the interpolated learning.

How may the present results be reconciled with those of the previous experiment by Kersh (1958), in which learning by discovery proved markedly superior? The preferred interpretation is that the findings of the two studies are actually complementary. Schematically, the treatments employed in the two experiments may be compared on a line representing the continuum of learning processes: at one extreme, learning without any external direction whatsoever (true self-discovery); at the other, learning by lecture-drill processes (rote learning), as follows:



As is indicated above, the Direct Reference treatment in the 1958 experiment is comparable to the Guided Discovery treatment in the present one; similarly, the Rule Given and Rote Learning groups correspond. The present experiment has no counterpart to the No Help treatment of the previous study; and, in the previous one, the Directed Learning treatment was not represented.

When compared as above, the results of the two experiments are remarkably similar. The initial achievement of the comparable

groups in both experiments was very high then dropped to where only about half of each group was able to recall and apply the rules after 4 to 6 weeks. In each experiment the differences in the performance of the Rote Learning and Directed Discovery groups was not notable; if anything, the Rote Learning groups tended to perform slightly better.

With respect to the motivating power of learning by discovery, in the 1958 experiment the superior performance of the No Help subjects on the retest together with their written comments and verbal reports to the experimenter strongly evidenced their increased interest. The present results leave no doubt that there is a tendency for interest to accrue as a result of learning by discovery.

The results of both experiments also are consistent in their failure to support the notion that attempts to provide added meaning will necessarily prolong memory for rules and procedures and will enhance their transfer. On the contrary, both experiments suggest that such attempts may well do more to interfere with learning than enhance it. This does not mean that rote learning is superior to learning with understanding. Rather it means that we need to know much more than we do about meaningful learning and how we come by it.

The relatively poor showing of the Directed Learning group in the present study is partially explained by the subjects' reported failure to practice the rules after the learning period to the extent that the subjects did in other groups. Why the Rote Learning treatment generated more interest than the treatment in question again may reflect nothing more than that the original learning was inhibited by the interpolated programmed learning. The subjects' unfamiliarity with the instructional procedure may have contributed to their confusion.

Most certainly the data from the two experiments under discussion suggest that the frequently taught principles of learning that pertain to self-discovery and meaning (see introduction) should be restated or qualified. The following statements are offered for further study.

Learning by self-discovery. Learning by self-discovery is superior to learning with external direction only insofar as it increases student motivation to pursue the learning task. If sufficiently moti-

vated, the student may then continue the learning process autonomously beyond the formal period of learning. As a result of his added experience, the learner may then raise his level of achievement, remember what he learned longer, and transfer it more effectively. The explanation for the elusive drive generated by independent discovery is not evident, but several have been offered, including the Zeigarnik effect of superior memory for unfinished tasks and the Ovsiankina effect of resumption of incomplete tasks. It also could be explained in terms of operant conditioning; specifically, as a kind of "searching behavior" reinforced by the experimenter's comments and by the subject's own successful progress toward a solution. Whatever the explanation, the motivating power evidently does not appear in strength unless the student is required to learn almost completely without help and expends intensive effort over a period of 15 minutes or more.

Meaningful learning. Aside from the advantage the student may come to have academically, he may not benefit from knowing the explanations for rules and procedures he learns, i.e., the pattern of relationships involved. That which is meaningful (understood) may or may not be retained longer and transferred more effectively than that which has been learned by rote. Moreover, superficial efforts to gain understanding after a rule or principle has been memorized may have an inhibitory effect when the student attempts to recall and transfer the original learning. If it is important only that the task be understood (as is most often the case, presumably), the essential relationships may be learned most economically when taught by another person or teaching program, not by process of self-discovery.

References

- AUSUBEL, D. P., Learning by discovery: Rationale and mystique. *Bull. Nat. Ass. Second. Sch. Principals*, 1961, 45, 18-58.
- BRUNER, J. S., The act of discovery. *Harv. educ. Rev.*, 1961, 31, 21-32.
- CRAIG, R. C., Directed versus independent discovery of established relations. *J. educ. Psychol.*, 1956, 47, 223-234.
- KERSH, B. Y., The adequacy of "meaning" as an explanation for the superiority of learning by independent discovery. *J. educ. Psychol.*, 1958, 49, 282-292.

- KITTELL, J. E., An experimental study of the effect of external direction during learning on transfer and retention of principles. *J. educ. Psychol.*, 1957, 48, 391-405.
- LI, J. C. R., *Introduction to Statistical Inference*. Ann Arbor, Mich.: Edwards, 1957.

IRVING MALTZMAN, SEYMORE SIMON,
DAVID RASKIN, and LEONARD LICHT
University of California, Los Angeles

Experimental Studies in the Training of Originality*

We rarely think we can make people creative and original. We usually believe that creative talent is an inborn quality. As teachers, we sometimes think that our job is only to stand out of the way of creative expression in students. If we can't facilitate it, we should at least avoid blocking it. We are forced to agree with Pope that "whatever is, is right": who are we to thwart Nature?

The words "creativity" and "originality" enjoy an exalted status in educational circles today. To question their position, if only to ask what observable behavior is creative and what is not, sometimes rivals the effect produced by questioning virtue. Given this climate of opinion, the investigations reported here are particularly important for two reasons: (1) originality is defined in terms of concrete and measurable behavior so that we know better what is being investigated, and (2) evidence is presented that certain conditions (which can be manipulated)

* Reprinted and abridged with the permission of the senior author and the American Psychological Association from the article of the same title. *Psychological Monographs*, No. 493 (1960), 1-23.

foster creativity—in other words, originality can be learned behavior.

The student should note that the investigators use a model of operant conditioning in their experimental design and in the explanation of results (see pp. 138–139). Since “originality” in other investigations is considered the highest form of mental activity, it is interesting to note that no distinction is made here between conditioning and training for originality.

In reading about these experiments the student should try to answer the following questions: (1) How do the authors define originality? Does this seem to be a satisfactory definition? (2) What techniques are used for obtaining original responses? (3) Which technique seemed most successful? Why? How could such a technique be employed in the classroom? (4) What evidence is presented in the report on Experiment V that originality is a form of learned behavior? (5) As a teacher, what would you do to promote originality as it is defined here? (6) Is originality the same as independent thinking?

A basic difficulty in attempting to facilitate original thinking is that it may not occur at all or at such infrequent intervals that reinforcements cannot be administered with sufficient frequency to effect an increase in such behavior. Thus, a fundamental problem in the training of originality is to devise methods for increasing its occurrence in the first place, thereby permitting the operation of reinforcement. We are assuming that originality can be learned and that the same principles of conditioning hold as in other forms of operant behavior. Some of the problems attendant upon this assumption have been discussed elsewhere (Maltzman, 1960). For our present purposes we need indicate only that by originality, or original thinking, we mean behavior that occurs relatively infrequently, is uncommon under given conditions, and is relevant to those conditions. In order, then, to facilitate the occurrence of original behavior, techniques must be devised for evoking many uncommon responses. Such training may then produce a disposition to give uncommon responses in other situations.

One procedure, used by Maltzman, Bogartz, and Breger (1958) is to repeatedly evoke different associations to the same stimulus words in a free association situation. This procedure prompts the subject to emit responses relatively low in the response hierarchy elicitable by each stimulus. They found that subjects receiving this training with intermittent reinforcement of uncommon responses were significantly more original on a new list of words than subjects without this training. A second experimental group which did not receive verbal reinforcement approached significance in comparison with the control group. When instructions to be original were added to the experimental and control treatments they showed a marked increase in the originality of their associations, and both experimental groups were now significantly more original than the control group. The Unusual Uses Test of originality devised by Guilford and his associates (1950) was administered following the free association test. Equivocal evidence for transfer of the association training effects to this situation was obtained in the form of a significant triple order interaction.

The essential features of the procedure responsible for the facilitation of original responses to new stimuli, however, cannot be determined from that study. The purpose of the present series of experiments, in part, is to determine what those relevant features are.

Experiment I

It may be that the important characteristic of the procedure employed by Maltzman, Bogartz, and Breger (1958) is not evocation of responses relatively low in the hierarchy of each stimulus word, as was assumed, but simply the production of many different responses. The purpose of this first experiment, therefore, was to determine whether facilitation of uncommon responses to new stimuli would occur when different responses were evoked by the presentation of different stimuli rather than by repetition of the same stimuli. Two different lists of stimulus words were employed for this purpose. One list consisted of relatively common and the other relatively uncommon words, in order to determine whether the frequency of usage of the words would have a differential effect upon the uncommon-

ness of test responses. Another control group not employed in the previous study (Maltzman, Bogartz, & Berger, 1958) was added to this experiment. This group responded in the same fashion to repeated presentations of a given stimulus word. Such a condition would enable us to determine whether the number of responses per se, or the number of different responses, is the relevant variable influencing test performance.

METHOD

Subjects. The subjects (Ss) were 292 students drawn from introductory psychology classes.

Stimulus materials. The stimulus words used in the initial list and the test list for all groups were selected from the norms obtained by Wilson (1942), and are the same as those used in a previous study (Maltzman, Bogartz, & Breger, 1958). Evocation of a relatively small number of different responses in a free association situation was the criterion for the selection of these stimulus words. The different training words presented to two of the groups following the initial list were selected from the Thorndike-Lorge count (1944). The list of relatively common words were chosen from among the 500 most frequently occurring words in the count. The list of relatively uncommon words were selected from words occurring not more than 6 times per 1,000,000. Guilford's Unusual Uses Test (1950) was administered after the completion of the free association test list.

Procedure. The Ss were treated in identical fashion on the initial presentation of the free association training list, the first 25 stimulus words for all the groups. At the start of the experiment the Ss received the usual free association instructions to respond as quickly as possible with the first word that came to mind.

After completion of the initial 25-word training list one control group (C) was given the test list of 25 new stimulus words. The previous instructions were repeated before presentation of the test list, as they were for all the other groups. A second control group (C_R) received five additional presentations of the same training list with instructions to try to give the same response to a given stimulus word each time. They received the test list following the last repetition of the initial training list.

Two experimental groups received a single presentation of the initial 25-word training list followed immediately by 125 different words. One group (X_L) received a list of words with a low frequency count, while the other group (X_H) received words with a high count. They were given the test list following the completion of the entire training list.

After completion of the training list a third experimental group (X) received five additional presentations of the list with instructions

to respond as quickly as possible to each word, but to give a different response than the one used before. The test list was administered following the last repetition of the training list.

All groups received the Unusual Uses Test immediately after the completion of the free association test list.

The experiment was conducted in group form. Stimulus words for the free association training and test were read aloud by *E*, and *S* recorded his responses in a booklet provided. Each page of the booklet had 25 numbered blank spaces for his responses. A stimulus word was presented every 5 sec. during the initial training list and the test list for all groups. The additional training words presented groups X_H and X_L were presented at the same rate. Repetitions of the initial training list words for groups X and C_R were presented at 10 sec. intervals.

RESULTS

The frequency with which the different responses occurred to each stimulus word was determined for the free association training and test lists. *Ss'* responses were then scored on the basis of the obtained frequencies, and each *S* was assigned an originality score which was the mean frequency of his responses. A low score therefore represents high originality. A few *Ss* failed to respond to a stimulus word or recorded an illegible response, but no *S* gave less than 24 readable responses. Their score was based on the number of legible responses recorded.

Table 1 shows the mean originality scores obtained by the various

Table 1
MEAN PRETEST FREE ASSOCIATION ORIGINALITY
SCORES IN EXP. I

Cond.	<i>N</i>	Mean	<i>SD</i>
C	64	91.69	19.74
C_R	62	92.70	16.42
X	60	86.39	25.89
X_H	60	97.06	19.35
X_L	46	95.89	20.64

groups on the first presentation of the training list, the pretest. A simple analysis of variance showed that they did not differ significantly prior to the introduction of the different treatments

($F = 2.22$; $df = 4,287$; $P > .05$). Homogeneity of variance was indicated by the Hartley test.

Table 2 shows the mean originality scores obtained by the various

Table 2
MEAN ORIGINALITY SCORES ON FREE ASSOCIATION
TEST IN EXP. I

Cond.	Mean	SD
C	76.38	23.18
C _R	70.90	20.99
X	54.49	18.42
X _H	53.40	24.57
X _L	56.44	18.64

groups on the free association test list. An analysis of covariance of the free association test scores was planned in order to increase the precision of the experiment. However, a test of the homogeneity of regression of the test on the training scores indicated that this assumption of the analysis could not be met ($P < .005$).

A simple analysis of variance of performance on the free association originality test was therefore conducted. An F of 14.95 was obtained ($df = 4,287$; $P < .001$). The results of t tests showed that the two control groups did not differ significantly from each other and the three experimental groups did not differ significantly. However, each of the three experimental groups was significantly more original than each of the control groups, $P < .001$.

Originality scores for performance on the Unusual Uses Test were obtained for each S by a procedure somewhat similar to the one employed for the free association test. The frequency with which different uses occurred for each of the six different objects mentioned in the test was first determined. In the previous study by Maltzman, Bogartz, and Breger (1958) an originality score was obtained for each S by counting the number of responses he gave to each test item that occurred less than 10% of the time in the obtained sample. This procedure was not feasible in the present study which employed more than twice the number of S s, since unique responses typically comprised more than 10% of the responses to each item. The originality score for each S therefore was taken as

his total number of unique responses. Any response was defined as unique if it occurred only once in the sample to a given test item. A second score obtained for each *S* was his total number of non-unique or common responses. This may be taken as a measure of fluency.

Table 3 shows the mean number of unique responses obtained by

Table 3
MEAN NUMBER OF UNIQUE USES IN EXP. I

Cond.	Mean	SD
C	2.58	2.30
C _R	1.03	1.06
X	4.75	3.54
X _H	2.27	2.08
X _L	2.72	2.36

the various groups on the Unusual Uses Test. The results from one *S* in Group C_R were discarded, because he incorrectly recorded his uses in the free association test booklet rather than the booklet provided for the Unusual Uses Test. Since Hartley's test indicated that the variances of these scores were heterogeneous ($P < .01$) and the distributions were highly skewed, it was decided to adopt the .01 level for rejection of the null hypothesis. An analysis of covariance was not conducted because of the absence of a significant correlation between unique responses and originality on the free association pretest.

A simple analysis of variance of the originality measure obtained from the Unusual Uses Test gave an F of 18.74 ($df = 4, 286$; $P < .001$). Since the variances were significantly different, half the appropriate degrees of freedom were used in the t tests of the differences between groups. Results of these tests showed that Group X gave significantly more unique responses than any of the other groups, $P < .001$. Groups X_H, X_L, and C did not differ significantly. However, Groups C and X_L were significantly more original than Group C_R at the .001 level. Group X_H was significantly more original than C_R at the .01 level. A rank order analysis of variance yielded comparable results ($H = 52.00$; $df = 4$; $P < .001$).

Table 4

MEAN NUMBER OF COMMON USES IN EXP. I

Cond.	Mean	SD
C	14.44	4.84
C _R	14.59	4.58
X	15.65	5.09
X _H	15.42	4.18
X _L	15.62	4.34

Table 4 shows the mean number of nonunique or common responses obtained by the different groups on the Unusual Uses Test. Hartley's test indicated that the variances were homogeneous. A simple analysis of variance yielded an F of .92, indicating that the groups did not differ significantly in terms of their fluency of common responses on the Unusual Uses Test. Analysis of covariance was again contraindicated by the absence of a significant correlation between frequency of common uses and originality on the free association pretest.

A basic assumption in the use of the training procedure employed with Group X is that it would increase the uncommonness of the responses to the training words. Fig. 1 shows the mean percentage of unique responses occurring on successive repetitions of the training list. It was necessary to employ the percentage of unique responses rather than the mean or median weighted frequency scores, because of the highly skewed distributions of scores obtained for the 25 training words. By the third repetition over half the responses of the average S were unique. It is clear from Fig. 1 that the training method employed with Group X had the desired effect of inducing an increase in the uncommonness of the responses emitted.

Group C and Group X in the present experiment are similar to conditions employed in the study by Maltzman, Bogartz, and Breger (1958). In the latter experiment the difference in originality between these two groups only approached significance on the free association test, but was significant when combined with verbal reinforcement or instructions to be original. In the present experiment Group X showed a highly significant increase in free association originality in the absence of verbal reinforcement or differential instructions to be original. In the previous experiment only

equivocal evidence of facilitation on the Unusual Uses Test was obtained. Here, the training administered Group X in the free association situation produced a highly significant facilitation of originality on the Unusual Uses Test.

The different results obtained in the two experiments may only be due to the increase in the size of the sample in the present experiment. However, it is possible that the change from an individual to a group experiment also contributed to the difference in results. Conducting a free association experiment in group form grants a degree of anonymity to S's responses which is not the case when he is run on an individual basis. In the group experiment

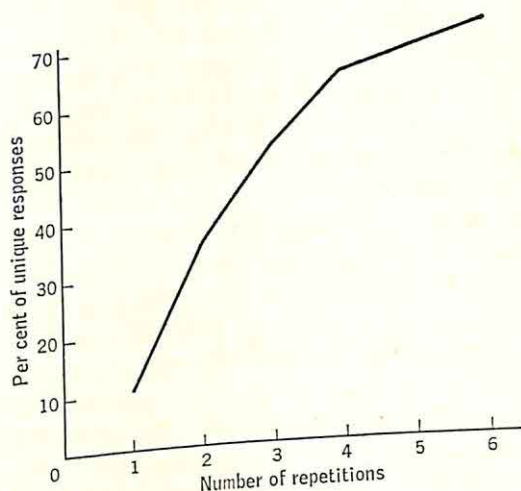


Fig. 1. Mean Percentage of Unique Responses Obtained by Group X with Free Association Training in Exp. 1.

there also are necessarily more different cues present that are potentially the occasion for many responses less likely to occur when S is alone with E in a bare room. Both of these characteristics may have contributed to the appearance of a training effect in this experiment of a greater magnitude than the one reported in the earlier study (Maltzman, Bogartz, & Breger, 1958). In any case the magnitude of the training effect obtained in this study is consistent with the results of the subsequent experiments reported here.

Groups X_H and X_L were employed in order to determine whether evoking many different responses would produce the same effect on test performance as the training provided Group X. The assump-

tion was that the occurrence of any given response increases the disposition of responses associated with it to occur. Evidence of such an effect has been reported by Judson and Cofer (1956) and by Judson, Cofer, and Gelfand (1956). Providing for the occurrence of many different responses during free association training should therefore increase the reaction potential of many more different responses on the test list as compared with the control groups. That an effect of this sort may be operating is suggested by the results of the free association test where the three experimental groups do not differ significantly, but are significantly superior to the control groups.

However, the results of the Unusual Uses Test in terms of originality scores indicates that the training administered Group X involves more than the arousal of many intraverbal associations. On this test, Group X gave significantly more unique responses than Groups X_H and X_L as well as the two control groups. These results indicate that the procedure of repeated evocation of different responses to the same stimulus words produces a more general disposition to produce uncommon responses than the procedure of evoking different responses to different stimuli. The effect of this training, furthermore, was restricted to increasing the number of uncommon uses. There were no significant differences among the groups in terms of their fluency of common responses.

The significantly poorer performance on the free association originality test of Group C_R than the three experimental groups indicates that the increased originality of the experimental groups on this test cannot be attributed to the number of responses evoked per se or to habituation factors. The three experimental groups and Group C_R emitted the same number of responses during training. The difference obtained must be due to the uncommonness of responses per se emitted by the experimental groups or the uncommonness of responses emitted to the given stimulus words.

The fact that Group C_R gave significantly fewer unique uses than Group C indicates that the training in the free association situation of repeatedly evoking the same response to a given stimulus interferes with the tendency to give uncommon uses. To some extent, the repeated occurrence of common responses to the same stimuli increases the tendency to emit common responses in other situations,

just as the repeated occurrence of uncommon responses to the same stimuli increases the disposition for original behavior in other situations.

Experiment II

The purpose of this next experiment was to further explore possible relevant variables and alternative methods of facilitating originality. Two groups of Ss were exposed to uncommon associations which they simply read, in order to determine whether emission of uncommon responses by S is a necessary condition of the originality training procedure. In an attempt to explore the extent to which the method of facilitating originality is peculiar to the training materials employed, a group of Ss was given repeated presentations of items from the Unusual Uses Test and tested with the free association materials. Conditions X and C of the previous experiment, which we shall hereafter call the standard experimental training and control conditions, were also employed in this experiment.

METHOD

Subjects. The Ss were 251 students drawn from introductory psychology classes.

Stimulus materials. The stimulus words used in the initial training and subsequent test lists are the same as those in Experiment I. However, their function was reversed in the present experiment. The list previously employed as the training list was used as the test list in this experiment, and the test list in the previous experiment was used as the training list. Guilford's Unusual Uses Test was administered after the completion of the free association test list to all groups except X₂.

Procedure. All Ss were presented the initial training list and the final test list of stimulus words in the same manner, including the usual free association instructions to respond as quickly as possible with the first word that came to mind.

The standard control group (C) received only a single presentation of the training list followed by the test list of words. The standard experimental condition (X₁), as in the previous study, received five additional presentations of the training list with instructions to respond as quickly as possible, but to give a different response than the one used before to each stimulus.

A second experimental group (X₂) was also asked to respond differently to repeated presentations. However, the stimuli employed

were the six items from the Unusual Uses Test. The Ss were instructed to write a different use following the oral presentations of each item. Five repetitions of the items were administered. A third experimental group (X_3) was presented with a booklet of 125 pairs of items. The Ss in this condition were instructed to underline the member of each pair which goes more readily with the stimulus word orally presented to them. The stimulus words were five repetitions of the training list. Response pairs were unique responses to these words culled from the results of Exp. I.

The final experimental group, X_4 , received the same response pairs as Group X_3 , but without the preceding stimulus words. These Ss were instructed to underline the member of each pair that they thought to be more familiar.

Stimulus words for the free association training and test lists were read aloud by *E*, and *S* recorded his responses in a booklet provided by *E*. The Unusual Uses Test was presented in written form to all groups except Group X_2 . Stimulus words were presented every 5 sec. during the initial free association training and test. Repeated presentations of the stimuli for originality training occurred at 10 sec. intervals. Fifteen sec. were permitted for the response to each unusual uses training item in Group X_2 .

RESULTS

Frequency distributions of the responses to each of the training words in Exp. I were employed as norms for scoring the responses to these words which were employed as the test list in the present study. The frequency with which the different responses occurred to each of the training words in the present experiment was determined, and Ss' responses were scored on the basis of these frequencies.

Table 5 shows the mean originality scores obtained by the various

Table 5
MEAN PRETEST FREE ASSOCIATION ORIGINALITY
SCORES IN EXP. II

Cond.	<i>N</i>	Mean	<i>SD</i>
X_1	64	87.72	20.02
X_2	46	83.56	21.37
X_3	48	87.85	18.36
X_4	39	85.66	18.98
C	54	83.48	23.18

groups on the first presentation of the training list. A simple analysis of variance indicated that the groups did not differ significantly prior to the introduction of the experimental treatments ($F = .55$). Homogeneity of variance was indicated by the Hartley test.

Table 6 shows the mean originality scores obtained by the vari-

Table 6
UNADJUSTED AND ADJUSTED (MEAN') FREE ASSOCIATION TEST
ORIGINALITY SCORES IN EXP. II

Cond.	Mean	SD	Mean'
X ₁	55.09	20.44	53.73
X ₂	80.02	25.88	81.54
X ₃	87.82	19.58	86.38
X ₄	84.37	21.90	84.43
C	80.67	24.08	82.23

ous groups on the free association test. Appropriate tests indicated that the variances of the test scores were homogeneous as well as the regression of test scores on training scores. The regression of the test scores on the initial training score was .69.

A simple analysis of covariance was therefore conducted in which the originality test scores were adjusted for initial differences on the training list. An F of 31.86 was obtained ($df = 4, 245$; $P < .001$). The results of t tests indicated that Group X₁ differed significantly from each of the other groups ($P < .001$), whereas these four groups did not differ significantly from each other.

The frequency with which different uses were offered for each item in the Unusual Uses Test was determined from the norms of Exp. I, and an originality score assigned each S based on the number of unique responses he submitted. A second score obtained for each S was his total number of nonunique or common responses. Since the score distributions for unique responses were skewed, and the variances were heterogeneous, as in the previous study, the transformation $\sqrt{X + .5}$ was applied to them.

Table 7 shows the mean number of transformed unique responses obtained by the various groups receiving the Unusual Uses Test.

Table 7

MEAN NUMBER OF TRANSFORMED ($\sqrt{X + .5}$)
UNIQUE USES IN EXP. II

Cond.	Mean	SD
X ₁	1.65	.56
X ₃	1.35	.58
X ₄	1.23	.54
C	1.17	.43

The variances of the transformed scores were homogeneous as well as the regression of unique responses on the free association pretraining scores. An analysis of covariance of the unique responses regressed on the pretraining scores yielded an F of 10.31 ($df = 3, 207$; $P < .001$). However, since the regression was only .02 no increase in precision over a simple analysis of variance was obtained, and the adjusted means which differ only slightly from the unadjusted means will not be presented. Individual comparisons among the groups by means of t tests showed that the standard experimental training condition, Group X₁, was significantly more original than each of the other conditions ($P < .001$). The other conditions did not differ significantly from each other.

Since the distributions of scores were highly skewed due to the large number of failures to give unique responses, the transformation could not normalize the data. A rank order analysis of variance was therefore conducted. Results comparable to the previous analysis were obtained. The H was 29.82 ($df = 3$; $P < .001$).

Table 8 shows the mean number of common uses obtained by the various groups on the Unusual Uses Test. The variances were homogeneous as well as the regression of common responses on the free association pretraining scores. An analysis of covariance of the

Table 8

MEAN NUMBER OF COMMON USES IN EXP. II

Cond.	Mean	SD
X ₁	20.09	5.17
X ₃	19.26	5.91
X ₄	21.08	5.63
C	17.03	5.18

common responses regressed on the pretraining scores yielded an F of 5.41 ($df = 3, 207$; $P < .001$). However, since the regression was only $-.02$ no increase in precision over a simple analysis of variance was obtained. Individual comparisons among the conditions by means of t tests showed that the three experimental groups did not differ significantly from each other. However, each of these three groups gave significantly more common responses than the standard control group ($P < .05$).

Results obtained in this experiment for Condition X_1 confirm the results of Exp. I that repeated evocation of different responses to free association stimuli is followed by an increase in the originality of responses to new stimulus words and to items on an Unusual Uses Test as compared to a control group (C) in which such training is omitted. Conditions X_3 and X_4 were introduced in order to determine whether the uncommon responses must be evoked by the stimulus words in order for the training to be effective. In the former condition uncommon responses were evoked as textual responses (Skinner, 1957) following the stimulus words, whereas in Condition X_4 only the textual responses occurred. Neither condition was followed by significant increases in originality on the tests employed as compared with Group C, and they obtained significantly lower scores than the standard experimental condition, Group X_1 . These results show that with the stimulus materials employed, uncommon responses must be evoked as intraverbal responses by the stimuli if there is to be an increase in originality in the test situations.

It may be argued perhaps that the associations provided Groups X_3 and X_4 are not uncommon. Although this is a possibility it seems rather unlikely, since each association occurred only once in a similar sample of approximately the same number of Ss. Furthermore, since a pair of uncommon responses was given for each stimulus word or ordinal position, twice as many uncommon responses were available to Groups X_3 and X_4 than to X_1 . Thus, even though some of these associations may have been common to the former groups they still should have received at least as many uncommon associations as Ss in Group X_1 .

Another hypothesis which may be suggested to account for the difference among the experimental conditions is that the procedure employed with Group X_1 induces the inhibition of common re-

sponses as the result of repeated presentations of the same stimulus words and the accompanying instructions, thereby permitting the appearance of uncommon responses. Thus, the facilitation of uncommon responses in the test situations occurs at least in part because of a disposition to inhibit common responses as well as an increase in the excitatory potential of uncommon responses. Conditions X_3 and X_4 do not provide for the inhibition of common responses, but evoke the uncommon responses directly as textual responses.

Evidence apparently contradicting this interpretation stems from another study (Maltzman, Brooks, Bogartz, & Summers, 1958) which found that uncommon uses occurring as textual responses may facilitate performance on the Maier two-string problem. Examination of the latter experiment suggests, however, that the results obtained there are not necessarily in conflict with the inhibition interpretation. In the latter experiment the uncommon textual responses were evoked in the presence of stimulus objects subsequently found in the two-string problem. The excitatory potential of uncommon responses to the test stimuli could therefore be increased by the training procedure employed. But in the present experiments the training and test stimuli were different. This suggests that a training method evoking uncommon intraverbal responses will produce a more general effect than training with uncommon textual responses, and implies further that if uncommon uses for objects not found in the two-string problem were to be employed, facilitation of problem solutions would not occur.

Although Conditions X_3 and X_4 did not show a significant increase in originality in the two test situations, they did yield a significant increase in common responses on the Unusual Uses Test, an effect also obtained in Condition X_1 . This finding suggests that uncommon responses, whether textual or intraverbal, tend to facilitate the occurrence of other responses, to facilitate fluency. However, since this result was not anticipated, the experiment was not designed with an adequate control group to determine whether an increase in common uses is peculiar to the evocation of uncommon responses during training or whether such an increase would occur as well if common responses were initially evoked. It should be noted that Condition X_1 in Exp. I did not show a significant in-

crease in common responses as compared to the control group, while highly reliable increases in original responses were obtained. This discrepancy in the results of the two experiments suggests that the present findings of increases in common responses should be accepted with caution. The results of subsequent experiments are equally inconsistent.

Experiment V

If, as we have assumed, originality can be learned according to the principles of instrumental conditioning, then it should show some degree of persistence following training. Since retention is one of the characteristics of learned behavior, the disposition to emit uncommon responses induced by the standard experimental training procedure should also show some degree of permanence. The purpose of the present study was to determine whether this in fact is the case.

METHOD

Subjects. The Ss were 177 students drawn from introductory psychology classes.

Stimulus materials. The lists of stimulus words used in the initial free association training and subsequent tests are the same as those employed in the first three studies, and were administered in the same order as in Exp. II. As before, the Unusual Uses Test was administered after the free association test list.

Procedure. All Ss were presented the initial free association training and final test lists of stimulus words in the same fashion. Two control groups received only a single presentation of the training list. Group C received the free association and Unusual Uses tests approximately one hour later. During the interval of time between the training and test sessions, Ss were engaged in work on a series of 10 paper and pencil problems. Group C₂ was dismissed following the initial presentation of the free association training list with instructions to return in 48 hours. The free association and Unusual Uses tests were administered at that time. Two experimental conditions, X and X₂, experienced comparable delays between the initial and final test sessions. However, as in previous experiments, they received five additional presentations of the initial free association list with instructions to give a different response than the one used before to each stimulus word.

RESULTS

Table 17 shows the mean originality scores obtained by the vari-

Table 17

MEAN PRETEST FREE ASSOCIATION ORIGINALITY SCORES IN EXP. V

Cond.	<i>N</i> ^a	Mean	<i>SD</i>
X	62	84.76	22.95
C	58	80.59	26.07
X ₂	27	77.46	21.50
C ₂	30	79.47	27.93

^a Due to an unfortunate oversight, approximately half the Ss in the two-hour delay groups were given a different form of the Unusual Uses Test which had the same frontispiece as the test previously used. We have therefore discarded the free association data from these Ss as well as their unusual uses.

ous groups on the initial presentation of the training list. A simple analysis of variance indicated that the groups did not differ significantly prior to the introduction of the experimental treatments ($F = .67$). Homogeneity of variance was indicated by the Hartley test.

Table 18 shows the mean originality scores obtained by the vari-

Table 18

UNADJUSTED AND ADJUSTED (MEAN') FREE ASSOCIATION TEST
ORIGINALITY SCORES IN EXP. V

Cond.	Mean	<i>SD</i>	Mean'
X	71.58	26.16	68.77
C	84.47	24.57	84.46
X ₂	77.12	23.13	79.20
C ₂	87.05	26.05	87.79

ous groups on the free association test of originality. Appropriate statistical tests indicated that the variances of the test scores were homogeneous as well as the regression of test scores on training scores.

Since the training and test scores were correlated significantly, a two-way analysis of covariance was conducted in which the originality test scores were adjusted for initial differences on the training list.

An F of 21.9 ($df = 1,172$; $P < .001$) was obtained for the originality training effects. An F of 5.58 ($df = 1,172$; $P < .05$) was obtained for the delay effect.

The frequency with which different uses were offered for each item in the Unusual Uses Test was determined on the basis of the norms from Exp. I, and originality and common response scores obtained as in the previous studies. Since the distributions of scores for unique responses were skewed, and the variances were heterogeneous, the transformation $\sqrt{X + .5}$ was applied to them.

Table 19 shows the mean number of transformed unique re-

Table 19

MEAN NUMBER OF TRANSFORMED ($\sqrt{X + .5}$) UNIQUE USES IN EXP. V		
Cond.	Mean	SD
X	1.42	.64
C	1.06	.45
X ₂	1.49	.53
C ₂	1.21	.57

sponses obtained by the four groups. Hartley's test indicated that the variances of the transformed scores were still somewhat heterogeneous, $P = .05$. As in previous studies, the measures of unique responses and common responses were not correlated significantly with originality on the initial free association list, thereby precluding the use of the analysis of covariance.

A two-way analysis of variance of the transformed unique responses yielded an F of 15.47 ($df = 1,172$; $P < .001$) for the originality training effect. An insignificant F of 1.10 was obtained for the delay effect.

Table 20 shows the mean number of common responses obtained by the four groups on the Unusual Uses Test. The variances were homogeneous. An F of .77 was obtained for the originality training effect, and 3.32 for the delay effect ($df = 1,172$, in both cases). Neither effect was statistically significant. The results from the two different tests of originality indicate that highly reliable originality effects do persist for some time. Thus, some additional support is gained for the interpretation of the training effects as a

Table 20

MEAN NUMBER OF COMMON USES IN EXP. V

Cond.	Mean	SD
X	18.06	4.84
C	18.15	5.24
X ₂	17.74	5.37
C ₂	15.67	3.97

form of learning. Results from the free association originality test also show a significant delay effect or decrement in originality as the time between the initial and final tests is increased from one hour to two days. This, of course, is another common characteristic of learned behavior.

Results from the Unusual Uses Test showed a highly reliable training effect but not a delay effect. There was no significant decrease in the number of unique responses when the delay was increased from one hour to two days. Again, this discrepancy in the results obtained from the two originality tests may be due to a variety of reasons. It cannot be concluded, however, that a reliable decrement in unique uses did not occur within the time interval studied. Comparison of the delay conditions employed in this experiment with the standard experimental and control conditions of Exp. II, which received the free association lists in the same order, suggests that a decrement occurred within the first hour of delay.

The extent to which the effects of originality training persist, or are retained, is undoubtedly a function of many different variables of the sort that have been studied in classical experiments on verbal retention. Such variables as amount of practice or original learning, a wide range of delay intervals, activity during the delays, and distribution of practice trials, may all influence the extent to which the effects of originality training persists.

As data accumulate from parametric studies of this sort, and they show that performance on tests of originality vary in a fashion similar to performance in more traditional learning situations, to that extent it may be assumed that the effects of originality training do not differ fundamentally from other forms of learning. The study reported here is an initial attempt in that direction, and lends some support to the hypothesis that originality can be learned in the same fashion as other forms of operant behavior.

Conclusions

Each of the four experiments employing the standard experimental training procedure with free association materials produced a highly reliable increase in the uncommonness of responses on two different tests. The first two experiments suggest that the standard experimental training procedure of repeatedly evoking different responses to the same stimuli is the most successful of the three experimental procedures employed. It produced significantly greater facilitation of originality on one or both originality tests as compared with the method of evoking different responses by presenting different stimuli as in Exp. I, or evoking uncommon responses as textual responses as in Exp. II. Possible variables responsible for the efficacy of the standard training procedure have already been indicated earlier in this paper and elsewhere (Maltzman, 1960).

The results of Exp. . . . V lend some support to the hypothesis that originality is learned behavior and varies as a function of the same antecedent conditions as other forms of operant behavior. Effects of originality training may persist for as long as two days, under the training and test conditions employed in Exp. V. Originality, at least, on the free association test, also varies as a function of the number of repetitions of the training list.

Although the standard training procedure has proved to be effective, there are many possible variations of the procedure that remain to be investigated. For example, instead of completing the entire training list of stimulus words before re-presenting it, successive responses to each stimulus might be evoked. Various combinations of these procedures may also be studied. Further experimental studies are also needed employing different types of test criteria, including the more traditional kinds of problem situations. Studies of this kind are currently in progress in the UCLA laboratory. Work of this sort may indicate the extent to which the kind of training procedure studied here might have practical applicability. In this connection, it should be noted that the behavior measured on the originality tests employed in these experiments and similar tests correlates reliably with ratings of original behavior (Drevdahl, 1956).

In conclusion, it should be noted that the term "originality" as employed in this paper refers to a particular kind of behavior measured by specified operations under given conditions. If it is objected that this is not genuine originality, then it should be incumbent upon the critic to specify the latter behavior in equally operational terms so that it too may be subjected to experimental study.

References

- DREYDAHL, J. E., Factors of importance for creativity. *J. Clin. Psychol.*, 1956, 12, 21-26.
- GUILFORD, J. P., Creativity. *Amer. Psychologist*, 1950, 5, 444-454.
- JUDSON, A. J., & COFER, C. N., Reasoning as an associative process: I. "Direction" in a simple verbal problem. *Psychol. Rep.*, 1956, 2, 469-476.
- JUDSON, A. J., COFER, C. N., & GELFAND, S., Reasoning as an associative process: II. "Direction" in problem solving as a function of prior reinforcement of relevant responses. *Psychol. Rep.*, 1956, 2, 501-507.
- MALTZMAN, I., On the training of originality. *Psychol. Rev.*, 1960, 67, in press.
- MALTZMAN, I., BOGARTZ, W., & BREGER, L., A procedure for increasing word association originality and its transfer effects. *J. Exp. Psychol.*, 1958, 56, 392-398.
- MALTZMAN, I., BROOKS, L. O., BOGARTZ, W., & SUMMERS, S. S., The facilitation of problem solving by prior exposure to uncommon responses. *J. Exp. Psychol.*, 1958, 56, 399-406.
- SKINNER, B. F., *Verbal Behavior*. New York: Appleton-Century-Crofts, 1957.
- THORNDIKE, E. L., & LORGE, I., *The Teacher's Word Book of 30,000 Words*. New York: Bureau of Publications, Teachers College, 1944.
- WILSON, D. P., An extension and evaluation of association word lists. Unpublished doctoral dissertation. Univer. of Southern California, 1942.
- WOODWORTH, R. S., & WELLS, F. L., Association tests. *Psychol. Monogr.*, 1911, 13, (5, Whole No. 57).

Appendix

THE CLASSIFICATION OF RESPONSES ON THE UNUSUAL USES TEST*

The list of common and unique uses that follows was initially obtained from the results of Exp. I and was used with minor modifications in the

* Classifications for "key," "safety pin," "watch," and "button" have been omitted.

scoring of the Unusual Uses Test from the later experiments. In the initial classification based on the results of Exp. I any use occurring only once in the distribution for all items was designated as unique. Uses occurring more than once were classified as common. Adjustments were made in the initial list by the addition of unique uses when they occurred in the test of a later experiment. On the other hand a unique use was reclassified as common if it occurred more than once on any given subsequent test. Few adjustments of this sort were necessary.

1. *An automobile tire* (used on the wheel of an automobile)

Common Uses

float	inner tube (for swimming)	make sandals
roll	train dogs	make money
swing	melt and reshape	life saving
bumper	boat dock (bumper)	to swim with
toy	wheel on cart	binding to hold tree
smoke	slingshots	horse equipment in horse
chair	start fire (burn)	show
obstacle course	put in garage	play catch with
jump on	blockade	draw circles
scrap rubber	football practices	buoy
flower bed (planter)	insulation	tunnel
weight	rubber bands	hold things in lining (hid-
mending (patch) rubber	throw things through	ing place, paper, water,
weapon (hit someone with,	(football, etc.)	etc.)
etc.)	sled	to throw
decoration (garden, gas	hoop	to cut up
station, etc.)	drum head	make smog
shoe soles	supporting another object	sand box
target	sell for living (retail prod-	to break things
advertisement	uct)	hide narcotics
costume	make smell	shock absorber
ride on		

Unique Uses

prop up bigger object	keep rug from slipping	make wall
holding hub cap	hold air	base for bell-ringing pole
protect car from bumping	test coordination	at carnival
paint	skin diver's net	trade it
protect axle	hold letters or bundles	base for lamp
lawn mower wheels	seat at beach	cover
toss bean bags through	press papers	fort
seat at ball game	picture frame	door mats
have on fishing boat	run in to build up legs	anchor down poles
cradle	soap box racer brake	durability demonstration
household rubber material	plug a hole	cable holder
use in musical chairs	in a skit	battering ram

Unique Uses (Cont.)

padding on walls of amusement park	hold back loose dirt	make rubber piece for trolling (fishing)
in relay games	keep tire salesmen in business	keep repair men busy
lift to build muscles	costume for strippers	roof for stork nest
trap	throw knives	souvenir
rubber strings for tying wire	makeshift table	ballast in a boat
hang on tree	gift	joke
clutter up backyard	slice and make holders	make mudpies in middle of
for color scheme	transfer air to bicycle	play tag with
push things down driveway	sit on while being pulled by speedboat	tear up and use as brake on toys
make tracks in sand	wagon	dogs to play with
use in circus act	to step on when there's mud	build fence
fill with cement for weight mold	to hide behind	end of hammer to pound out dents
capture or hold someone	chewing gum (Spain)	keep wood raft afloat
build pyramid to jump off	stick pins in	tractors
draw on with chalk	steal	flatten something
build tree house	wheel barrow	little boy to play army in
sunbaths	in a clock	burn to prevent frost damage
texture pieces in art projects	adjust temperature on a safe	store unused garden hose
loan to your neighbor	rubber coverings	ash tray
put in a continental kit	support walnut trays above ground	bird nest
hold back flood waters	in zoo for monkey to hang on	make an invention
build with pillow	around a horse in a bull fight	pile up outside of dealers
line doors or furniture	inner lining for rubber gun	cause accidents
store refuse in	hang things on	protective lining for mechanical clamp
bumper cars at beach	wedge	bottom of table legs to prevent harming floor
bed for animal	carve on	
hold door open		
painter's still life		

Pace of Presentation, Number of Trials, and Amount of Practice Time as Determiners of Learning*

The question of the number of trials and the distribution of practice needed in learning a given amount of material has been a traditional topic of investigation by learning psychologists. In earlier readings in this book the question has been raised several times, by Gagné and Bolles (pp. 47-48), by Duncan (pp. 219-220), and by Maltzman and his associates (pp. 287-307).

In the late nineteenth century Hermann Ebbinghaus found in an investigation of memory that he could learn a list of non-sense syllables more efficiently if he spread (distributed) his practice over three days rather than concentrating (massing) it in one day. Thus the issue of mass versus distributed practice was born. Since that time knowledgeable teachers have tried to persuade students of the benefits of regular (distributed) study over cramming (massed) study. For reasons of their own (not necessarily related to efficient learning) students have been ignoring this advice. Many consider practice to be "busy work," and others find that they do quite well on their teachers' tests while still massing their practice.

The report of Kopstein seriously questions the beneficial effects of distributed practice. It also questions another, related principle which makes the number of trials a crucial learning variable. All that remains of the original theory is *amount of practice*. Kopstein's results merely suggest that learning is

* Published with the permission of the author. Footnotes, tables, are omitted.

learning—that if something is learned the first time through, further trials and practice may not be necessary. Perhaps learning is *less* doing than Dewey imagined.

The following questions are suggested: (1) How do the results of the second experiment compare with the first? (2) Why is the issue of *covert* processes raised? (3) How could the time which is now devoted to practice in the classroom be more fruitfully used for original learning?

In an earlier and as yet unpublished experiment (Kopstein, 1960) it had been found that Ss who themselves determined the length of the stimulus inspection period and of the response anticipation period learned a list of paired-associates in a significantly lesser number of trials than Ss for whom these periods were set (quite brief) by E. However, S-paced and E-paced groups required the same amount of time to achieve the criterion of one correct list anticipation. These results, although they confound stimulus presentation and response anticipation periods, suggested the possibility that neither the pacing of the trials (in effect the distribution of practice), nor the number of trials themselves had had much effect on the mastery of the list. Because it runs counter to commonly held beliefs, this possibility was investigated; also, other suggestive evidence had been obtained before (Wright and Taylor, 1949; Kopstein and Roshal, 1961).

Experiment I

This experiment was conducted to investigate several ancillary matters as well as obtaining preliminary evidence concerning the Pacing-Trials-Time issue. Only the immediately relevant portions of it are reported here.

METHOD

Design. Four experimental groups were designated. Groups I and II and for practical purposes Groups III and IV practiced for the same total length of time. Groups II and III practiced for the

same number of trials. Groups I and III and Groups II and IV practiced at the same rate (time/trial).

Learning Materials. In order to maintain comparable conditions the list used earlier (Kopstein, 1960) was retained. It was composed of these six paired-associates items:

Red:	PEOS	MIFA	SPEROUT	CIF
Green:	USFAS	KETT	SIMRO	CUZ
Yellow:	NORRPI	TATVIN	SECMI	JEPA
Blue:	LARP	PRIT	CANULC	MAIT
White:	JOAP	NOLM	PUTI	NOMIPA
Black:	SPERI	VEM	MOGI	SAKO

Each color name stimulus term was associated with a four part response term. The pseudo-words constituting the response terms of this list had been manufactured from randomly chosen words falling among the first one thousand in the Thorndike-Lorge count (1944). For each vowel and consonant in the original word another had been substituted. Thus relative vowel-consonant frequencies (though not their sequential dependencies), and normal vowel-consonant relationships of common English were preserved. All items were typed with an IBM electric typewriter using a carbon ribbon and were then photographed as 35 mm. slides. Eight random sequences of six were established from a table of random numbers and eight sets of six slides were so arranged.

Apparatus. Slides were projected by means of Spindler and Saupe Selectroslide 35 mm. automatic projector. Its magazine accommodated 48 slides; when eight trials had been completed the entire cycle was repeated. The projection period was controlled by a Tandem Recycling Timer, Type A manufactured by Industrial Timer Corp. Responses were entered in special response booklets. Each page contained one stimulus term and four blanks for the four parts of the response term.

Subjects. The Ss were twenty-seven male and fifty-three female college students. They volunteered their services, but were paid for their efforts. Ss reported in groups of five at a time and were randomly assigned to their experimental treatment as a group. Although this does not constitute an entirely random assignment of Ss to treatments and may include a Type G error (Lindquist, 1953),

for this preliminary experiment the risk associated with it was ignored.

Procedure. After Ss were seated appropriate tape recorded instructions for the initial training period were read to them. The instructions were the same for all treatments. Then the projector and the timer controlling the slide changes were started. *E* kept track of the elapsed time and at the end of the scheduled period he turned off the projector and the timer. Experimental Group I practiced for 30 minutes at the rate of 7.5 seconds per item for a total of 40 trials. Experimental Group II practiced for 30 minutes at the rate of 30 seconds per item for a total of 10 trials. Experimental Group III practiced for $7\frac{1}{2}$ minutes at the rate of 7.5 seconds for a total of 10 trials, and Group IV practiced for 9 minutes at the rate of 30 seconds per item for a total of three trials. The inequality in the actual practice time between Groups III and IV was necessitated by the fact that, if Group IV had practiced for only $7\frac{1}{2}$ minutes, only $2\frac{1}{2}$ trials would have been completed. Probable effects of the extra $1\frac{1}{2}$ minute of practice were thought to be negligible for preliminary purposes.

At the end of the training period test booklets were passed out and instructions for taking the test were read. Fifteen seconds were allowed for responding to each item. At the end of the fifteenth second *E* called out "turn" and simultaneously reset his stop-watch which was started again when most Ss had successfully turned to the next page. As soon as the last item had been completed the test booklets were collected and Ss were dismissed with the caution not to discuss the events in the experimental session.

RESULTS

The Ss' performance was measured by the number of correct responses given on each item. Only those responses were scored as correct which appeared in conjunction with their proper stimulus term and also in their proper position among the response term parts. An analysis of variance showed mean scores of the four experimental groups to differ significantly ($F = 13.470$, $p < .001$). Since inspection of the raw scores showed their distribution to deviate substantially from normality, they were subjected to an arc-sin transformation.

However, although it induced normality, the net effect of this transformation was merely a very slight increase in the F value.

An analysis of the differences among Group means by Duncan's New Multiple Range Test (Duncan, 1955) produced the results shown in Table 1 [omitted]. Duncan's test permits the assessment of the mutual differences that may exist among a set of obtained means. Since the calculation of Duncan's original tables involved certain approximations, Harter's corrected tables published by Edwards (1960) were used.

It will be seen that the mean for Group II is numerically slightly larger than the mean for Group I, but that the difference is not significant. Both of these means are substantially larger than means for Groups III and IV, and these differences are also significant beyond the .01 level of confidence. Again, the mean of Group III is slightly larger than that for Group IV, but a statistically significant difference is not demonstrable.

DISCUSSION

The outcome of this preliminary experiment supports the expectation that practice time was the most important factor in determining the degree to which the list was mastered. Number of trials did not appear to be a factor, since mean scores of Groups II and III (same number of trials) differed significantly. Similarly, distributional effects appear to be absent, since mean scores of Groups I and III, and of Groups II and IV (same duration per trial) differed significantly from each other. At the same time the null hypothesis could not be rejected for the observed differences in mean scores between Groups I and II and Groups III and IV (same length of practice).

Experiment II

This experiment was undertaken to extend and confirm the findings of Experiment I. Three aspects of that experiment cast some doubt upon its reliability. First, the findings rested upon a comparison of only two data points and nothing could be said about the extremes of the continua involved. Second, it was thought that

unequal difficulty of the learning materials (pseudo-words) that were employed might have given rise to a statistical artifact. Third, because each treatment had been run as a group, the possibility of an artifact due to the confounding of the group-factor with the experimental treatment could not be entirely dismissed.

METHOD

Design. This experiment, like the previous one, sought to pit the factors of "Number of Trials," "Total Practice Time," and "Pacing" (Rate of Presentation per item) against each other. A simple 4×3 factorial design was chosen. The first factor, "Pacing" (or "Presentation Rate," or "Distribution of Practice") represented a 4 second, 8 second, 16 second, and 32 second presentation time per item. This is equivalent to respective paces of 24 seconds, 48 seconds, 96 seconds, and 192 seconds per trial (presentation of complete list). The second factor, "Total Practice Time," allowed either 6 minutes 24 seconds (384"), or 12 minutes 48 seconds (768"), or 25 minutes 36 seconds (1536") for the learning of the experimental list.

The successive levels of both of the factors result from a doubling of the value for the preceding level. As a consequence the relationship of cells along any upper-left-to-lower-right diagonal is such that the same number of trials occurs albeit in different total amounts of practice time. By the simple addition of two extra cells (2 seconds per item for 384 seconds practice, and 64 seconds per item for 1536 seconds practice) a second 4×3 factorial design could be made to emerge which pitted a "Trial" factor and a "Practice Time" factor against each other.

Ten subjects were assigned to each of the 14 cells in this experiment for a total of 140. However, each of the two separate (though overlapping) designs represented by it, had only 12 cells with 120 subjects; 10 of the cells were common to both. Each of the two distinct designs had two cells which it did not share with the other one. In every other respect their features were identical.

Learning Materials. Except for the changes noted here the learning materials in the present experiment were identical with those in Experiment I. The change made was to eliminate the pseudo-words and to substitute for them a set of standard nonsense syllables.

These nonsense syllables were drawn randomly from Glaze's 53% association value list as reproduced in Stevens (1951, p. 542) and then were successively substituted for each of the 24 pseudo-words. The revised list was as follows:

Red:	FAP	ZAM	TIQ	NUV
Green:	VOK	XAN	KER	GAC
Yellow:	VOM	FOV	LER	QAT
Blue:	NUR	CEG	RIK	WEP
White:	SOH	JIR	QIX	VAY
Black:	ROF	BOQ	JED	BIJ

The six items in the list were photographed on a 35 mm. slide-film in 10 randomly ordered cycles (trials).

Apparatus. The slide-film was fashioned into an endless loop and projected by means of an SVE Soundview projector. Changeover from frame to frame was, for practical purposes, instantaneous and was triggered by two Lafayette timers operating in an tandem cycle. The closing of the timer-switch, which triggered the frame advance on the SVE projector, also caused a 60 watt incandescent bulb to flash. A solenoid-actuated Veeder-Root counter was connected in parallel with the light bulb as a double check on the accuracy of the timers, since the total number of separate images to be projected under each treatment could be calculated.

Subjects. 140 male and female college students served as Ss. As before these Ss volunteered their services, but were paid. However, in this experiment Ss reported individually and assignment to treatments was entirely random.

Procedure. Upon reporting Ss were given a test booklet (identical with those used in Experiment I) and asked to write their name on the front cover. Ss were cautioned not to open the booklets until told to do so. Next a tape recorder played back the pre-recorded instructions which were essentially identical with those used in Experiment I. Thereafter lights were turned off and projection of the experimental materials began. An image of approximately 8 in. $\times$ 12 in. was projected over the S's shoulder onto a screen approximately 5 ft. from the S's face. Immediately above the area on which the image appeared a socket was mounted which held the 60 watt in-

candescent lamp mentioned above. With each image (item) change this lamp flashed briefly. After the last item had appeared the projector and timer were turned off, lights were turned on again, and the tape recorder started to give the instructions for the test. Instructions and procedure for the test were essentially identical with those used in Experiment I except that the time allowed for each response was 12 seconds rather than 15 seconds.

RESULTS

Performance was scored as in Experiment I.

"Pacing" vs. "Practice Time." Table 2 [omitted] shows that only "Practice Time" achieved an F -ratio that is convertible to a p value well beyond the .001 level. Neither "Pacing," nor the interaction between "Practice Time" and "Pacing" yielded appreciable F values. An arc-sin transformation of the scores had only a negligible effect on F values.

The actual values of the mean number of correct responses are tabulated in Table 3 [omitted]. It may be seen that mean values along the rows in the table which represent the different presentation rates are of the same order of magnitude. By contrast the columns representing the different practice times differ sharply from each other.

It is clearly apparent that treatment means which involve the same pacing are widely separated while those representing treatment means with the same total amount of practice time fall essentially on a straight line.

"Trials" vs. "Practice Time." Table 4 [omitted] shows that in this case, too, only "Practice Time" attained an F value at an exceedingly high confidence level. Neither "Trials," nor the interaction between "Practice Time" and "Trials" generated an F value worthy of note. An arc-sin transformation failed to alter the pattern of results.

The actual mean number of correct responses obtained with the different treatments are tabulated in Table 5 [omitted]. Again it may be seen clearly that mean values from row to row in the table representing the different numbers of trials are of the same order

of magnitude. On the other hand, columns representing the different practice times contain values that are of decidedly different magnitudes.

DISCUSSION

Doubts remaining at the conclusion of Experiment I concerning the true determiner of how much is learned are assuaged by the outcome of Experiment II. The pattern of results strikingly illustrates the influence which length of practice time has on the mean learning achievement. In Tables 3 and 5 means of treatments having a common length of practice time can be connected by lines that remain well within the error-range surrounding a straight horizontal line. (A slight qualification of this statement will be discussed below). Thus, not only did a progressive change in pacing (the rate of presentation) fail to produce any change in slope but, more amazingly, a progressive increase in the number of trials failed to produce the familiar "learning curve." Finally, since the separation between all corresponding means in the three columns is essentially the same in Tables 3 and 5, it is clear that no interactions were involved. As in Experiment I, the distribution of practice (Trials $\times$ Practice Time) must be ruled out as a significant factor.

In arriving at the experimental design for Experiment II and in choosing the values which the several levels of the factors involved should take, an attempt was made to include the extremes. For example, it had been empirically determined that two seconds were the minimum required for completely reading any one of the items in the experimental list just once. Thus a two second presentation rate represents the meaningful upper limit on "Pacing." However, because of the nature of the experimental design, this treatment mean is not given in Table 3. It does appear in Table 5 as the entry for the 32 trials and the 384 seconds practice time. That this treatment did not produce results different from others involving the same amount of practice time was borne out by Duncan's New Multiple Range Test (Duncan, 1955).

Another extreme occurred in the case of the treatment which specified a presentation rate of 64 seconds per item. This treatment, too, is not given in Table 3, but appears in Table 5 as the entry for

four trials in 1536 seconds (25 minutes 36 seconds). Unlike the upper extreme (2 second rate) this lower extreme (64 second rate) does seem to have had a depressant effect on performance. Duncan's test shows that the mean for this treatment does not differ either from the means of the group of treatments involving 1536 seconds of practice time, nor from those involving the 768 seconds of practice time. This effect may be a "boredom effect." It was observed during the conduct of the experiment that Ss receiving this treatment would become restless after prolonged stimulus exposure and would look away from the display.

Especially when taken together, the results of the two experiments clearly indicate that a simple count of the observed number of trials (usually regarded as equivalent to number of reinforcements) has serious defects as an experimental measure. Moreover, the distribution of trials within some practice period (the pace of practice) could not be shown to be of significance. This finding has been corroborated by several investigators (Briggs, Plashinski & Jones, 1955; Briggs, 1961; Silverman, 1961). Hence, sheer distribution of practice, when imposed by *E* and when unconfounded with intra-trial and inter-trial rest periods, must be rejected as a meaningful concept.

The crucial point in this matter is the issue of *overt* rather than *covert* occurrence. Since only practice time as an independent variable functioned as a significant determiner of mastery, but since time per se cannot be regarded as a causal factor, it must be assumed that covert processes took place over time. This then suggests that such classic findings as, for example, Hovland's (1938) must be reinterpreted, since (a) the duration of covert learning activity is minimally controlled, if at all, by *E*, and (b) the insertion of "rest pauses" either within, or between lists cannot be assumed to terminate completely all covert learning activities of the Ss.

Summary

Two parallel learning experiments were reported which pitted pace of presentation, number of trials, and amount of practice time against each other. The learning task was to master a list of six paired-associates each having a common color name as stimulus term

and a four part response term consisting of pseudo-words or of nonsense syllables. Only the total duration of practice influenced the degree of mastery over the list that was obtained. Neither pacing of trials or distribution of practice, nor the number of trials had any significant effect. The findings are interpreted to mean that the Ss' learning activity was covert and beyond control of *E*. Hence, observed number of trials is rejected as a meaningful experimental measure; the traditional conception of distribution of practice effects is questioned.

References

- BRIGGS, L. J., Self-pacing versus automatic pacing in a paired-associate learning condition. In A. A. Lumsdaine, ed., *Student Response in Programmed Instruction*. Washington, D.C.: National Academy of Sciences, National Research Council, 1961 (in press).
- BRIGGS, L. J., PLASHINSKI, D., & JONES, D. L., *Self-pacing versus automatic pacing of practice on the subject matter trainer*. Laboratory Note, ASPRL-LN-55-8, Air Force Personnel and Training Research Center, 1955.
- DUNCAN, D. B., Multiple range and multiple *F* tests. *Biometrics*, 1955, 11, 1-42.
- EDWARDS, A. L., *Experimental Design in Psychological Research*, 2d ed. New York: Holt, Rinehart, and Winston, Inc., 1960.
- HOVLAND, C. I., Experimental studies in rote-learning theory: III. Distribution of practice with varying speeds of syllable presentation. *J. exp. Psychol.*, 1938, 23, 172-190.
- KOPSTEIN, F. F., An experimental exploration of the interaction between teaching and learning behavior. Unpublished Ph.D. dissertation, University of Illinois, 1960.
- KOPSTEIN, F. F., & ROSHAL, S. M., Verbal learning efficiency as a function of the manipulation of representational response processes. In A. A. Lumsdaine, ed., *Student Response in Programmed Instruction*. Washington, D.C.: National Academy of Sciences, National Research Council, 1961 (in press).
- LINDQUIST, E. F., *Design and Analysis of Experiments in Psychology and Education*. Boston: Houghton Mifflin Company, 1953.
- SILVERMAN, R. E., *Response mode, pacing and motivational effects in teaching machines*. U.S. Navy Training Devices Center, Technical Report NAVTRADEVCEEN 507-3, 1961.

- STEVENS, S. S. ed., *Handbook of Experimental Psychology*. New York: John Wiley & Sons, Inc., 1951.
- THORNDIKE, E. L., & LORGE, I., *The Teacher's Word Book of 30,000 Words*. New York: Bureau of Publications, Teacher's College, Columbia Univ., 1944.
- WRIGHT, S., & TAYLOR, D. W., Distributed practice in verbal learning and the maturation hypothesis. *J. exp. Psychol.*, 1949, 39, 527-531.

Introduction

In experimental psychology the subject of communication and language is often obscured because rats cannot talk and the college student, as an experimental subject, is often asked *not* to talk. Yet the most significant aspect of human behavior may be the use of words. Surely, barring an educational revolution in the schools, words will remain our chief means of communication. Therefore the educational psychologist is justifiably concerned with verbal communication, especially as it occurs in the classroom.

Communication is defined by Lorge as the "process by which an individual transmits stimuli to another to modify the receiver's behavior" (p. 327). The breadth of this definition permits this chapter to include readings on topics as diverse as the measurement of meaningfulness, verbal logic, and reading. The definition also suggests a problem facing American education. Simply stated, it is that a single message (stimulus) must be understood (in terms of the sender) by an audience of receivers who are infinitely various in themselves and who, in addition, are influenced by the variable social and cultural contexts within which the message is received. In these circumstances the effects of all communication become less predictable and controllable. For example, in the privacy of our homes a parental grunt may speak volumes to the children who receive the message. In a larger social and cultural context the same grunt may be considered only bad manners. The teacher, especially in the American urban school, is a unique creator of messages (stimuli) which he sends to a motley student body, receivers who are products of the American genetic, cultural, and national melting

pot. Also, now that our educators and teachers have taken to the stage and the studio to produce television programs and films, the educational audience grows more rapidly in size and variety than our knowledge of the effects of these communications on it. Although the next chapter is devoted to the use of the mass media and audio-visual aids in the schools, we can observe here that television and films are only part of the problem. The use of tape recordings, phonograph records, films, and video-tapes only adds to our communication problems in education.

The concept of communication as defined here ought to remind teachers that the messages they send should not only be received but also understood. They are understood when the behavior of the receivers (that is, the students) is changed in the way the teacher desires. Frequently teachers do not make any systematic effort to discover the effects of their verbal communication, even in the face-to-face relationship of the classroom; although the same ambiguous and misleading assignments are given to students term after term, poor student performance is not traced to what Lorge calls communication *noise*. Lectures are rarely tested to see if they have conveyed the ideas as the lecturer intended or whether some accidental characteristic of the communication caused students to misinterpret what was spoken. Test questions, which are written communications between the instructor and student, are seldom examined for possible ambiguity, misleading cues, needless redundancy, and non-readability. Even something as obvious as the effects of the teacher's general style of speaking and the accompanying gestures and posture, perennial subjects for student satire and mimicry, are unknown to classroom teachers of long experience. Textbooks are selected without careful examination of the treatment of content for clarity and effectiveness of communication. There is much talk in the American school, as there must be in any school, by both teacher and student, but there is often little knowledge of how well it contributes to achieving the objectives of the curriculum.

Relationship of Readings in Chapter 5

Only one reading in this chapter is a research report, although all the readings indicate lines of research in communication which could prove helpful to the schools. They all discuss the uses of

language and therefore are closely related to discussions in the previous chapter, which was concerned largely with verbal thinking. Lorge's discussion of communication in a psychological framework contains definitions and concepts, such as redundancy, context, and readability, which have proved useful in educational research on communications and language. Reading remains the major communication skill taught and used in our schools. In the selection by Roger Brown the controversy concerning the phonetic or look-and-say method of teaching reading is reviewed against the background of psychological and linguistic theory and research. The reading by Othanel Smith is a rather severe reminder that the study of thinking is not the exclusive province of the psychologist and that thinking and language still belong in the school. The last reading, by Miller and Selfridge, is a redefinition of the concept of meaningfulness.

IRVING LORGE

Late Professor of Education

Teachers College, Columbia University

How the Psychologist Views Communication *

Psychology is surely not the only approach to the study of language and communication. Rhetoric, semantics, the study of cognition, linguistics, and logic are also fruitful approaches. The psychological study of language and communication, however, highlights aspects of behavior not as clearly conceptualized in the other fields.

In the area of communication, Lorge discusses the high redundancy of our native tongue, which increases the likelihood that messages will be accurately received. Redundancy is a kind of repetition which attempts to guarantee certain changes

* Reprinted with the permission of Mrs. Irving Lorge and the publisher from the article of the same title, *Teachers College Record*, 57 (1955), 72-79.

of behavior in the receiver. Teachers have long been aware of the need for repetition as practice or review to consolidate learning gains. Repetition as redundancy in communication would be an area for experimentation, for example, in giving directions to students, in the demonstrations of skills, and in the explanation of principles. Redundancy would also seem to be a characteristic of programmed learning (pp. 136-208), which attempts to supply enough verbal prompting to prevent a mistaken response. The student should note how the ideas of redundancy and of learning context are used by Brown in his discussion of the dispute over reading (pp. 338-355). Learning context as expected word sequences is used by Lorge as an argument to support the "look-and-say" method for the teaching of reading.

Lorge's review of the research on language learning may also have implications for teaching: (1) For example, if, as some investigators have claimed, within the first few months of the infant's life he makes all the sounds of every language ever spoken—even the Mongolian dialects—what implication may this have for foreign language training in the schools? (2) If the reinforcement of certain sounds but not of others is the explanation for specific forms of language development, would aural programming in the foreign languages be the way to regain the lost sounds? (3) If language is the vehicle for cultural indoctrination, how can we avoid cultural parochialism in our language teaching? (4) Does the advice on the use of word counts agree with what Ausubel has written about the use of the concept of readiness in the schools?

The verb "to communicate" has many meanings. Its Latin roots carried the significations of dividing and sharing, of making common to many, and of imparting. These senses survive in the current meanings of to transmit, to impart information, and to share and enjoy in common. Thus, the dictionary now distinguishes among the three aspects of communication so as to separate process, message, and effects. These aspects cannot be separated except for emphasis

in exposition. The psychologist brings them together by his definition of communication as the process by which an individual transmits stimuli to another to modify the receiver's behavior. Such a definition implies that the communicator may be his own receiver. The communicator may modify his own behavior in and by the very act of talking to, or writing for, himself as well as for others.

Communication thus involves the reciprocal interactions of sending and receiving signals, of composing and understanding messages, and of sharing and enjoying ideas. These three interactions may be likened to interrelated stages involving the areas of engineering, psychology, and sociology. The engineering aspect deals with the means by which signs are sent and received accurately regardless of their meaning. The psychological emphases are concerned with acquisition of language in its variety of meanings. The social level deals with consequences of interchanges of communication.

Code as Communication

In the transmission of messages in telegraphic code, many of the psychological problems in communication are revealed dramatically. What are the difficulties in learning to send and receive code, or what are the possible confusions between different code signals? Keller, in a succinct historical review of the development of Morse Code brings some of the problems into clear focus. He reports that "Samuel Morse designed his telegraph to record *visual*, rather than auditory, code. . . . The first Morse code was cryptographic. There were numbers for all of the letters and many of the words of the English language. Morse spent many hours building a 'telegraphic dictionary' to be used in coding and decoding messages. These hours were wasted. His dictionary was barely completed when he decided to use an 'alphabetic' code in which each letter had its own signal."¹

The visual Morse code very quickly gave way to the now well-known auditory code, because telegraphers found it more practical to listen to the *sounds* the instrument was making than to view the graphic record. Indeed, operators became so adept that they were

¹ Fred S. Keller, "Stimulus Discrimination and Morse Code Learning." *Transactions of the New York Academy of Sciences, Series II. Volume 15* (April, 1953), pp. 195-203.

able to identify the individual who was sending the code, by his "touch" or sound pattern.

The transition from visual to auditory code indicates that adults found it easier to get the message by hearing than by seeing. Essentially, the experiences of the early telegraphers suggest that accuracy in receiving messages can be increased when the proper sensory modality is utilized. When a person suffers from a sensory handicap, it is necessary to substitute different means for reception and transmission of communication, such as tactile reception of Braille patterns by the blind, or visual reception of manual signs and gestures by the deaf.

Morse, in developing his graphic code, originally had planned to use a sequence of from one to ten dots to represent the digits 1, 2, 3, 4, 5, 6, 7, 8, 9 and 0. He realized, however, that adults cannot judge quickly and accurately at a glance the differences among five, six, seven or more dots. Accordingly, he adopted the dots, dashes, and intra-signal spaces which now constitute American and International Morse. The kind of discrimination that the individual had to make necessarily limited the nature of the pattern that could be used.

Still another communication problem is illustrated in the reception of American Morse code. Keller reports that a defect of the code was the possible confusion of five code signals: 6, P, H, 4, and 8. Unless 4, 6, and 8 occur in a context of numbers, 4 may be called V, 6 may be called P, and 8 may be called B. As for the letters when alone or in cipher, P may be called H, and H may be called S. In the absence of meaningful context, the signals may be misheard or misunderstood.

Of course, in the absence of context, any external noise or any momentary lapse of attention by the receiver may lead to a misunderstanding of the signals and a distortion of the message. Engineers try to develop mechanisms that will minimize the errors in reception by reducing the amount of irrelevant noise in the channel of communication. Examples of such disturbing noise would be static in radio reception and engine noises in airplanes and submarines. In reading, the analogy of noise would be the lack of adequate contrast between the print and the page, or an excessively small type size for the reader.

Telegraphic code may be sent as a cryptic message or in the clear. Cryptic code is used to conceal the message from all but the intended receiver. Usually, cryptic code is some very specialized arrangement of letters which when referred to a code book can be deciphered to arrive at the proper meaning. For instance, the five-letter sequence P Z H A V may be a cryptic message which in the clear means "Happy birthday and many good wishes for the years to come." If the message were sent in American Morse, the receiver might mishear the code and record it as H Z S A V. Obviously, there would be no internal evidence to suggest that the H should have been P or the S should have been H. On the other hand, if the message had been sent in the clear, so that the receiver knew that all words transmitted were to be English, he could have utilized his learned knowledge of English to check the reception. For example, if the telegrapher had been recording, "Plans are beginning to take . . ." and heard as the next word "S H A H E," it is quite likely that he would write the word "SHAPE" either consciously on the assumption that the sender had made an error, or unconsciously because of the familiarity of the whole message. The probability that the word had to be SHAPE may have been so great that the receiver wrote it as it *should* be.

In sending cryptic code, or in sending messages in a background of noise, there would be a marked advantage in transmitting the message several times. In the event of noise or accidental errors in sending, it is unlikely that the same noises or errors would occur at the same places in the message. An additional protection would be to have the receiver retransmit to the sender the message he had received. Examples of such repetitions are confirming an order, or calling back the directions in landing an aircraft. Whenever a message is composed of elements for which the receiver has little foreknowledge or expectancy, the repetition of the message decreases the probability of error. For the code message P Z H A V the receiver has no foreknowledge or expectancy that the letter following P is likely to be Z or that the letter after A must be V. In English, however, if SHA were received, certain letters would be quite unlikely as the sequel and others would be very probable. Further, the whole message increases the probability that the fourth letter will be P. Some shorthand systems, indeed, are based on the concept

that it is not necessary to record all the letters of a word. For instance, in the Phillips Code which was developed to speed the telegraphic transmission of press reports, the word *acquit* is abbreviated as *aqt*, *acquitted* as *aqtd* and *acquitting* as *aqtg*. The letter *u* in *acquit* is redundant in that it does not add information to the transmission of *acquit*. In English, *q* is always followed by *u*. Actually, the *c* does not really add very much either.

Redundancy in the sending of a message refers to an overdetermination of its reception. Redundancy can be increased by repetition of the message either by the sender or by the receiver, or redundancy exists because the receiver has learned the probabilities in messages. The probabilities can be within words, as in expecting that the next letter after TH is E; or it can be among words in meaningful sequences, as in the message "Plans are beginning to take . . ." for which the learned sequel is in the idiom "to take shape." Redundancy, whether learned by experience or planned as a technique to overcome noise, increases the probability of correct reception at the expense of increasing transmission-reception time or effort.

Learning Context

The advantage of learned redundancy is illustrated in the classical study by Bryan and Harter² in the way student telegraphers learned their trade. In one experiment, the amount of improvement in sending and receiving messages in the clear and in cryptic code was compared. Three conditions were contrasted: connected discourse, unrelated words, and letters in cryptic code. The rate of improvement was most rapid in connected discourse, somewhat less rapid in words, and markedly less rapid in independent letters of cryptic code. The difference is probably attributable to the fact that the young adults were able to utilize their previously learned knowledge of the sequence of letters in a word or of words in a sentence. The

² William Lowe Bryan and Noble Harter, "Learning a Life Occupation: The Telegraph Operator," in William Lowe Bryan, Ernest Hiram Lindley, and Noble Harter, *On the Psychology of Learning a Life Occupation*. Indiana University Publications, Science Series No. 11, 1941, pp. 69-129. (This is a reprint of the reports that originally appeared in the *Psychological Review*, Vol. 4, 1897; Vol. 6, 1899.)

greater improvement for connected discourse illustrates the potency of positive transfer from early to later learning. The three kinds of skill—letter, word, and discourse—all show improvement, but the rate of gain depends upon the previously over-learned language habits even though the student was not completely aware of the kind and amount of his previous acquisitions. It may be hypothesized that the plateaus Bryan and Harter found in learning to receive telegraphic code are attributable to the previously learned language patterns of letter sequences in words, word sequences in sentences, or idiomatic sequences in connected discourse. It is not so much that "a hierarchy of habits" develops in the course of practice as that previously acquired language habits can be transferred, in whole or in part, to the new task.

In connected discourse, context implies expected word sequences. Indeed, one definition of context may be in terms of conditional probability dependent upon previous occurrences. For instance, the probability that *Plans* will be followed by *are* may be less than the probability that *Plans are* will be followed by *beginning*. Hundreds upon hundreds of successful experiences with word sequences give the individual these probabilities even though he may be unaware of them. The probabilities that are learned depend upon the nature of the language to which he has been exposed. Each language has its own orthography, syntax, and idioms. How, then, are these patterns learned?

The new-born infant does not need to develop or invent the gestural, vocal, or written language. While many theories have been created to account for language, so far none has been substantiated. Whether language originated as a form of vocal imitation of gestures or as a response to strong emotional stimulation is not nearly so important as the fact that hundreds of different vernaculars exist. No child has ever learned his society's vernacular without the intervention of at least one other human who had knowledge of the language. Many people would like to believe that human infants could develop a language of their own by themselves. Occasionally, they suggest that the so-called "wolf children" create their own language for themselves. Exciting as such a demonstration would be, no case of a "wolf child" has ever been authenticated. Indeed, the evidence suggests that such a doctrine is not very probable. Gold-

farb's study of foundling children reared in the absence of the social stimulation of language play between parent-surrogate and child suggests that a human must give the infant the rudiments of language.³

Language Learning

The infant, at birth, does cry. It probably does not have available, however, the mechanisms for making all the phonetic sounds that constitute speech. Infants must develop not only in the ability to make sounds but also in the sensory modalities to receive signals. Auditory acuity, for instance, is not fully developed until about the twelfth year, and visual acuity somewhat later. Research does indicate, however, that within the first six to nine months, the child develops the ability to utter any sound or pitch variation necessary to speak any vernacular. Children reared by Chinese-speaking parents or by surrogates whose vernacular is Chinese learn Chinese speech. Children brought up by Bantu-speaking parents or their surrogates whose speech is Bantu learn that speech pattern. It is interesting to speculate that in the United States, four million newly born infants each year learn the language patterns from their parents and the other children or adults in the family unit. No two families provide the same training or experience, yet the vast majority of these infants will acquire not only the language pattern for making and receiving communications within the family but, more significantly, sufficient speech to be able to understand and be understood by others. The learning of language by the infant is, perhaps, the most beautiful demonstration of teaching and learning. In the course of learning the vernacular, some speech sounds will be reinforced or rewarded, and some speech sounds will not be reinforced. Within a very few years the child will be able to utter reinforced sound patterns but may not be able to make the sounds that were never or rarely rewarded. For instance, Chinese adults may find it difficult if not impossible to say the *l* sound, substituting for it the nearly equivalent *r* pattern.

³ William Goldfarb, "Infant Rearing and Problem Behavior," *American Journal of Orthopsychiatry*, 13, 1943, 249-265 (April); and "Variations in Adolescent Adjustment of Institutionally-Reared Children," *American Journal of Orthopsychiatry*, 17, 1947, 449-457 (July).

As the child learns the sound patterns that will bring him attention, comfort, and food, he acquires more than speech. He learns the names of people and things, the tonal pattern for demanding, asserting and questioning. But even more, as the child learns his vocabulary and tonal rhythms, he gets attitudes and values. He learns which sound patterns are acceptable enough to bring satisfaction and reward, and also those which the parents prohibit and taboo. Learning the vernacular is more than acquiring the words and syntax. Language learning by the young child means learning the culture in which he lives and grows. Learning language means learning to behave as a human in the society. Language, for the child, is a way of behaving. It is not, initially, a mode for thought or for reflection. Of course as the child matures he can use language in specialized ways. In early life, however, it is the means by which he gets control from the environment, and by the same token controls the social scene.

Psychologists have recognized the importance of language development. One of the measures of child learning is the number of words that a child knows in the sense of understands. Vocabulary probably increases steadily from the first year of life to beyond the sixtieth birthday. Vocabulary has been used as an index of intellectual development. Certainly it is an important measure of the rate of learning the culture. The child acquires during the lifelong process of learning his language the concepts as well as the cultural values.

In the early learning of language, it is the individual child who *individually* learns the cultural expectations. The family unit teaches the child its language, but the speech the child gets is social. It is the means by which any individual can (and usually does) interact with other individuals and with groups. Sociologically, the language the child acquires reflects not only the taboos but also the social sanctions. The very words developed within a society are indicators of the social goals as well as the significant features of the local environment. In any vernacular, word counts of the most frequent words and their most frequent meanings give a concrete measure of the more important aspects of the society.

By the time the child is five years old he has had the relation between words and values so strengthened that he has become conditioned to behave with certain probabilities to words and word

patterns. Sophisticated adults, too, will probably respond in certain ways to language. "I am now going to serve you a delicious charcoal-broiled steak, some lettuce and tomatoes, and a baked potato" may bring about salivation and certain anticipatory gustatory reactions in you. The extent of these reactions will depend upon the time you read the sentence, the amount of attention you give it, and the amount of belief you have that the promise will be fulfilled. Your behavior cannot be predicted with certainty but your probable behavior may be estimated in the sense that it can be ascertained what proportion of people would salivate or would feel pangs of hunger.

Understanding Communication

The effects of the sentence, moreover, depend on your understanding of it. What does *understand* mean? The word has many meanings. It includes such senses as "to get the idea of," "to apprehend by way of information," "to have information about," and "to accept as significant."

Prerequisite to understanding a message is getting the message from the sender to the receiver. After that, understanding of the message is a complex involving getting the meanings of words, perceiving the relations among words, appreciating the intention of the sender, evaluating the significance of the message.

The concept of *understanding* has been used in the development of reading comprehension tests. Less than fifty years ago, one of the primary objectives in the teaching of reading was to achieve expressive oral reading. Children were evaluated by the quality of the dramatic rendition of the printed text. All too frequently, children who *read* well, did not understand what they were reading. E. L. Thorndike's famous "Reading as Reasoning" (Thorndike)⁴ demonstrated that elementary school children were not very efficient in comprehending the texts that they were asked to read silently. Thorndike's articles, and the series of reading tests he and William A. McCall developed, changed significantly the goals and the evaluation of reading instruction. The reading comprehension tests

⁴ Edward L. Thorndike, "Reading as Reasoning; A Study of Mistakes in Paragraph Reading," *Journal of Educational Psychology*, 8, 1917, 323-332 (June).

emphasized the appraisal of understanding by asking children to answer questions that gave indication of getting the meanings of words in context, getting the main idea of the passage, getting specific facts from the text, and making inferences and predictions from the materials read.

Now, with television, radio, motion pictures and other audio-visual mediums so readily accessible, as much attention should be given to the evaluation of listening and viewing. Can children get the over-all significance of an illustration, appreciate details in it, see relations among objects? Can children see the intentions of the producer of a television show? Intensive studies of the utilization of the newer communication devices are needed. Understanding of these devices, too, may be evaluated by asking children to answer questions. In a sense, the questions that the children must answer may influence the manner in which the materials—text, speech, or picture—are understood. The questioning technique, in a genuine sense, influences the understanding of the communication.

Nevertheless, tests of silent reading have indicated some reasons for malcommunication. Among these, were the textbook writers' use of a vocabulary so difficult or erudite that children failed to get the meaning of the passage because they did not know some of the key words. Or again, the sentence structure was so unusual that the child could not extricate himself from the maze. For instance, consider a sentence like "Were it not that King Arthur was grievously hurt, he would have gone to the wars." Not only are there some difficult concepts, but there is also the unusual sentence structure which may cause the young reader to stumble on his way to understanding and appreciation.

Improving Communication

Thorndike and other educational psychologists raised the question "Are the words in children's books the right ones?" They answered the question by assuming that the most frequent words in print were the ones the child should learn to understand. Later, the concept of frequency was extended to measuring the usualness of sentence structure and, more recently, to codifying the most frequent words by the frequency with which each of its different mean-

ings was used in print. The frequency counts of words, of meanings, and of sentence structure significantly affected the preparation of textbooks, especially on the elementary school level. The psychologists provided the means for increasing the probability of comprehension of texts. Texts can be made more understandable by changing vocabulary, sentence structure, idea density, and idea sequence. About twenty-five years ago the first demonstration of the potentialities for improving communication by changing words and structure was made in the experimental evaluation of three contrived versions of the same content in an information bulletin for farmers. It was found that vocabulary with Anglo-Saxon roots, variation in sentence length, and logical organization from the known to the unknown would make a text not only easier to understand but also one the reader preferred to read.

The frequency counts and the empirical study of style elements were the forerunners of the so-called "readability formulas." These formulas were developed as a means for estimating the difficulty of texts for specified audiences such as school pupils and semi-literate adults. All readability formulas attempt to measure the over-all difficulty of a text in terms of internal structure. The items in readability formulas include one or more of the following: vocabulary difficulty, sentence length, idea density, and human interest. Vocabulary is measured by some indication of relative commonness of the words, such as that among the most frequent thousand or five thousand words in print, or by some estimation of relative difficulty, such as syllabic length. Sentence structure is appraised in terms of average sentence length. Empirical study has shown that the less usual sentence structures have the larger number of words per period. Idea density can be measured in terms of prepositional phrases, because phrases tend to be used to pack a text. Idea density, of course, can be estimated by the number of different concepts per hundred words. The greater the number of different nouns, for instance, in a hundred words, the greater the difficulty. Human interest is usually appraised by the relative number of personal nouns and pronouns, such as mother, father, you, and they. Each of these elements separately or in combination leads to an estimate of text difficulty. But important as these elements are, they do not evaluate idea sequence or organization. To this degree, then, readability

formulas do not measure all the elements that make for comprehensibility.

Mechanical manipulation of a text by selecting more frequent words, making shorter sentences, or eliminating prepositional phrases will not necessarily make a text easier. Changing "I am going to town" to "I am going townwards" does not make for greater comprehensibility. Readability formulas are not rules for writing; they are merely useful, but not completely efficient, yardsticks.

Comprehension is prerequisite to the effects of reading for enjoyment, evaluation, and action.

Lasswell has suggested a paradigm for studying public communication. Basically, the formula may be summarized as "Who says what to whom via what channel with what effect?" In the interchange between speaker and hearer or writer and reader, the comprehension of the communication involves not only understanding the content but also the nature of the communicator and his intentions. Social psychologists have been devoting an ever-increasing amount of attention to the individual and social consequents of communication. They are the historical successors to Aristotle in that they are attempting to evaluate the effects of communication. The rhetorician is concerned with how he got his effects. The psychologist would like to produce evidence about it. Variations in the text or in the speech will produce different effects in different individuals and in different groups. In general, differences in the speaker's or writer's gestures, use of aids, will not produce uniform effects in the listeners or readers. The very differences among the individuals in education, or in social and economic status, in motivations, and in emotional maturity will produce different results in the members of the reading and listening audience. Perhaps one of the greatest difficulties in social communication is that the communicator makes assumptions about the receiver's ability to understand the message. For instance, when he prepares a text, he may assume implicitly that if he knows the meaning of a word, or can comprehend a metaphor, or responds to an appeal, the listener or the reader, too, will understand, appreciate, and be motivated. Researches, however, now demonstrate the error of the implicit assumption about the nature of the audience. Certainly, except for two-way conversation, the best assumption about an audience is that

it is variable in ability, in knowledge, and in motivation. The effects of communication always have an element of uncertainty. Nevertheless, much has been learned about the relation among characteristics of the communication, the way it is delivered, and the effects upon the audience in terms of enjoyment and action.

The psychologist studies all aspects of communication at all levels. He is concerned with transmission and reception of messages, with learning and comprehending ideas, attitudes, and values, and with appraising the consequences of communication in comprehension, pleasure and action. His emphases vary from the application of knowledge for the improvement of communication to the development of new knowledge about it. The communications revolution in no small part reflects his contributions.

ROGER BROWN

Massachusetts Institute of Technology

A Dispute About Reading*

Whether or not, as Roger Brown suggests, the dispute over reading is partly the result of the guilty consciences of parents who (so they secretly believe) spend too much time before the television set and too little time reading, it can be flatly stated that few educational issues can arouse more verbal tirades in the popular press than the issue of how reading is taught. Even the researchers seem to lose their composure and abandon their scientific objectivity for polemic and debate. Yet all of this controversy makes a sober review of the evidence more urgent, and further research crucial. Paul Woodring shows how neither the wild charges made against the reading methods presently used by the schools nor the extravagant claims made by educators in defense of the schools come close to the truth. He

* Reprinted with the permission of the publisher, from Roger Brown, *Words and Things*, 1958, The Free Press of Glencoe, Inc., pp. 65-81.

writes, "By any reasonable definition of reading the great majority of boys and girls today can and do read. If 'Johnny' is taken to mean an average American child, the charge that 'Johnny can't read' is arrant nonsense. He can and does."[†] However, if the question is not whether Johnny reads better than his ancestors but rather does he read as well as he can and should, the answer is firmly, No.

The following discussion is in the editor's opinion a sober and relatively complete review of the dispute over which method of reading gets better results, the phonetic or the look-and-say method. It is partly a historical review of the problem and shows how the influence of psychological research and theory has combined with the educational philosophy of progressive education to establish the look-and-say method in many of our schools. As for the psychological contribution of this period, the student should note the limitations of the research of Cattell, particularly his assumptions about adult and child behavior in reading. He should see how Gestalt psychological theory, although originally used to support the look-and-say method with its emphasis on learning whole words and sentences and on the grasp of meaning, can be used to support either teaching method. Brown uses it here to argue for an "insightful phonetic method." Of course, the issue can only be settled by research which, one imagines, could have provided the solution in the first place. The student should also examine Brown's proposal for reading instruction to see in what ways, if any, it is consistent with Kersh's findings on directed versus independent discovery (pp. 277-287).

METHODS OF LEARNING TO READ

Imagine that you are teaching the primary grades in America and you will find even more cause than Shaw had to be angry about our writing system. Your pupils are first graders, varying somewhat in mental age but probably averaging six years. They have been speaking English for about four years. According to Smith, they are likely to have well over 10,000 words in their recognition vocabularies. You

[†] *Saturday Review*, "Can Johnny Read?" January 20, 1962, p. 40.

must teach them to read and then to spell. If our writing were consistently phonetic you could simply teach them the letters of the alphabet corresponding to each sound, give them a little practice in analyzing words into sound elements, and they would have reading vocabularies as large as their speaking vocabularies. Their 10,000+ words could be read with comprehension, or at any rate with all the comprehension attached to the spoken words. The child would spell out each new word, recognize the result as one of his 10,000 familiar speech forms and understand the written version as he understands its spoken equivalent. In fact, however, things are not so easy and this is clearly illustrated by the letter *A*. The name of that letter is *ay*. This is sometimes the sound of the letter in an actual word (as in *ate* or *ape*) but the letter is more often pronounced (as it is in *at* or *and*) as a short vowel. Which of these phonetic values should you teach? Even if you teach both there are horrible errors to be anticipated when your pupil finds *A* in *boat*, *peak* and *beauty*. As for *B*, the name of the letter (*bee*) begins with the most common phonetic value of the letter but also includes a vowel that is not ordinarily associated with the letter (as it is not in *but* or *bill*). Then, ought one to tell children about *doubt* and *debt* in which *B* has no sound? Some letters have names which do not even contain the sound most commonly associated with the letter. The sound of *H* is usually that heard in *he* but that sound is not contained in *aitch* the name of the letter. Neither is the sound of *W* in *double you*. The names of letters in the English alphabet are never the same as the sound most commonly associated with the letters and, furthermore, for most letters there is more than one common sound value. The English alphabet is so inconsistent in its phonetic values that it might be a good idea to teach the system as if it were not phonetic at all.

Phonetic training with the alphabet seems to work very well in European countries. In Germany and Italy children are said to become literate in their first two years of schooling. The spelling bee is not a popular contest in these countries for the reason that nearly everyone can spell most words. After World War II, the American occupation forces in Germany tried to introduce the spelling bee as part of the democratization program but they failed because of the uniformly high level of spelling prowess. Lessons in the phonetic values of letters work very well in Europe, but per-

haps that is because the languages involved have more consistent phonetic spellings than has English. A method well adapted to Italian is not necessarily well adapted to English. To be sure, reading has been taught by this method in America and England but the results in spelling accuracy and reading skills are not dazzling. It may be that there is a better method to use with English.

Look-and-Say. About thirty years ago the majority of American teachers decided there was a better method and they gave up or minimized the older alphabetic and phonetic methods. The new technique is called the look-and-say method. The fundamental idea is to treat each word as a unique visual pattern, rather as if our writing system were semantic with a different form for every meaning. The fact that these forms are constructed of a small set of recurrent letters is not stressed because the sound values of the letters are not constant. Writing is put in direct contact with meaning and its relation to speech is not taught because that relation has grown too ambiguous to be useful. Training begins with the short common words that the first grade child has long had in his speaking vocabulary. Characteristically each word is mounted on a card with a picture of the object named. The teacher flashes the card, pronounces the name repeatedly and calls attention to the picture. Vocabulary necessarily builds slowly since each word-referent association must be independently memorized. However, it is customary to begin with the words most frequently seen in printed English. From such a list of common serviceable words simple stories have been composed which a child can read while his vocabulary is still small.

The look-and-say method has often been described as a scientific method founded on psychological research as the old-fashioned phonetic methods were not. Perhaps the most often cited experiments are those done by Cattell in 1885. He showed by two lines of evidence that a familiar word is read as a whole rather than by spelling out its letters. In reaction time experiments the response of naming a short word is very nearly as quick as that of naming a single letter. This suggests that word recognition is a unitary act very much like letter recognition. Cattell also showed, using the gravity chronometer for quick exposures, that the time required to read letters that do not make words is about twice the time re-

quired to read letters that do make words. This result also suggests that the word is recognized as a whole pattern rather than as a combination of letters. Javal found in 1878 that an adult reader's eyes do not move steadily along a line, passing from letter to letter, but rather move saccadically, i.e., in a series of jumps. Erdmann and Dodge demonstrated that the fixation pauses are the times of effective exposure in reading. Evidently the adult reader recognizes a number of word shapes in a single glance. If this is the method of the accomplished reader why train the novice to analyze words into letters? These researches suggested to educators that adult reading techniques might be used from the first.

The look-and-say method also found justification in theoretical psychology. After the First World War Gestalt psychology began to influence American work. The Gestalt point of view, developed by Max Wertheimer and his students Kurt Koffka and Wolfgang Köhler, was more closely linked than associationistic behaviorism with highly developed forms of human perception and problem solving. Gestalt theorists stressed the importance of an overview of the whole as a prerequisite to the meaningful, intelligent solution of problems. They insisted that rote memorization of meaningless parts was not an important kind of human learning. Since the new method for teaching reading dealt with words, the meaningful wholes of which letters are meaningless parts, it seemed to be in accord with Gestalt theory.

Finally, the look-and-say method seemed better suited than phonetic drill to a new philosophy of education that held it more important for children to be happy and wise than for them to be well stocked with every kind of information. The ability to spell out words is not the most important aim of training in reading. What we need is more adults who read with interest and understanding and who seek out high quality reading matter. It seemed likely to many teachers that phonetic drill would cause children to develop an enduring distaste for reading. Stopping to spell out letters would slow them up, break the line of thought, leave them bored and inexperienced. Dealing with meaningful materials and whole stories from the beginning, a child trained by the look-and-say method would be more likely to understand what he read and to develop into an avid adult reader.

For all the reasons I have given most American teachers turned away from old-fashioned phonetic methods. But now, some thirty years after the change, they are being scolded for having made a frightful mistake. Their most censorious critic is probably Rudolf Flesch. His book, *Why Johnny Can't Read* was an American best seller for many weeks. Because the book has been a best seller, academic folk, educators and psychologists, have been inclined to ignore it or to depreciate it as a cheap effort to scandalize the public. Certainly Dr. Flesch presents his evidence like a prosecuting attorney. There is plenty of rhetoric and an occasional tendency to stack the cards. Generally, however, the argument is sound. So we will take Flesch seriously, but also study the teachers' rebuttal.

For and Against Phonetic Training. If you test reading prowess you find that it consists of many skills which are not necessarily in correlation. The particular skill in which America's Johnny is supposed to be deficient is the ability to sound out new words, to read aloud material he has not seen before. If each word is taught as a unique visual pattern it follows that one will only be able to read the words on which specific training has been given. These will not be very numerous for Johnny since it is common nowadays to set 1,300 words as a reading goal for the first three years of instruction. Some parents have found that Johnny reads well within this list but can do nothing at all with new words. This is distressing since he cannot very well have classroom training on all the words he will eventually need to read. Seashore and Eckerson state that adult recognition vocabularies run well over 100,000 words. At the rate of 400 new words a year, it will take Johnny 250 years to reach his parents' level. There was something in the old system of training that taught you how to read new words as well as the old ones used in your lessons.

Of course, if you are taught sound values for letters you can sound out new words a letter or two at a time and so do a reasonably good job of reading new material. But is this reading? The proponent of word recognition methods may see nothing useful in being able to *pronounce* new words that are not understood. The usefulness of being able to sound a new word depends on the state of the reader's speaking vocabulary. If the word that is unfamiliar in printed form is also unfamiliar in spoken form the reader who

can sound it out will not understand the word any better than the reader who cannot sound it. Even so the ability to pronounce the new word (however ineptly) has some advantages. If the word is encountered in private reading it can be carried by pronunciation to parent or teacher for definition. However, the real advantage in being able to sound a word that is unfamiliar in print, only appears when the word is familiar in speech. The child's letter-by-letter pronunciation, put together by spelling recipe, will, with the aid of context, call to mind the spoken form. There will be a click of recognition, pronunciation will smooth out, and meaning will transfer to the printed form. The ability to sound out new words is not simply a pronunciation skill; it is a technique for expanding reading comprehension vocabulary to the size of speaking comprehension vocabulary. This is a considerable help since speaking vocabulary is likely to be ten times the size of reading vocabulary for the primary school child.

It is not quite fair to say that the child trained to whole word recognition has no techniques that can be used for recognizing new words. He will have learned something about the probabilities with which words follow one another in English, something of the sequential probabilities of the language. Suppose a child who has learned to recognize words as unique patterns has never seen the word *lion* though he has often spoken it. If he comes upon that word in the sentence *The lion is in the zoo*, he can guess such probable insertions as *monkey* or *tiger* or, sometimes, *lion*. He cannot look to the word for help in choosing between these alternatives. Since the word is a totally new pattern there is nothing to be learned from it. The words he guesses (like *tiger* and *monkey*) may not look or sound at all like *lion*. They will resemble *lion* only in that they are sometimes found in the same sentence contexts. A child who has learned how to sound his letters will also have learned something about English sequential probabilities. In addition, however, he has a second set of cues—the rough sound of the word—to help him choose among possible alternatives. It is not that the child who recognizes whole words is without resource when faced with new words but rather that he has one less resource than the child who knows how to sound his letters. Much of our growth in reading vocabulary comes by working out unfamiliar words. Surely two

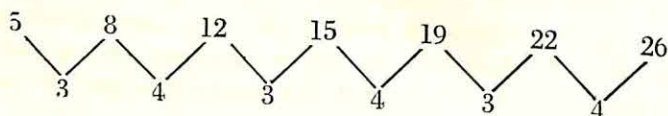
methods of word attack are better than one. Admittedly there are inconsistencies in English spelling but it remains a phonetic system with inconsistencies—not a semantic system.

What substance is there in the supposed experimental and theoretical support for the look-and-say method? Consider first the Cattell experiment showing that letters making words are read more rapidly than letters that do not make words. This result has been interpreted as a proof that adults read the "whole word picture" rather than individual letters. We can propose another interpretation. Printed English has a high level of redundancy. When every other letter of a running text has been deleted, a practiced reader of English will still be able to reconstruct most of the original (Shannon). Perhaps Cattell's subjects were able to read letters in words more rapidly than letters not in words because in the former case unobserved letters could be guessed from those that were identified while in the second case this was not possible. Letters in words follow sequential probabilities familiar to readers of English while letters at random are all equally probable at every juncture. It is quite possible therefore, that Cattell's subjects were reading individual letters rather than "total word pictures" and were able to report more letters than they could possibly identify at very brief exposures because the additional letters could be inferred from those observed. Reading research of the last fifty years indicates that while the general shape of a word has some cue value, the clear view of letters is a more important factor in word identification. Phonetically trained readers probably need to see all the letters in the beginning. As they store the sequential probabilities linking English letters fewer visual cues are needed. The adult reader is able to identify many words at a glance but it may be that this ability is best developed out of letter-by-letter reading.

Consider now the presumed support deriving from Gestalt theory. The words "wholist" and "meaningful" are very frequently used by Gestalt theorists to characterize their own position in contrast with that of the associationists. It is not surprising, therefore, that, on a cursory reading, the Gestalt work should seem to favor the reading method that deals in "meaningful" units—*whole* words. I think, nevertheless, that this is a misconstruction of Gestalt theory. Words differ from letters or other phonetic elements in that they have

reference to non-linguistic objects or events and, in this sense, words are "meaningful." In the writings of Gestalt psychologists, however, the word "meaningful" is more nearly synonymous with "systematic" than with "referential." "Meaningful" learning is learning that fits into a structure. Meaningful material is material in which there are systematic relations among the elements. Words are larger units than letters since words are compounded of letters and so, in a sense, words are "wholes." When Gestalt theory calls our attention to "wholes" it is to the system that determines the character of its parts. Are words the relevant "wholes" for someone learning to read?

Perhaps an example of a Gestaltist analysis of a learning problem will make clear what is meant by systematic, wholist learning. George Katona presented the following series of numbers to be learned: 5 8 12 15 19 22 26. Katona had subjects study the series until they could perfectly recall it. One week later he asked them to reproduce the series but no one could do so. With another group of subjects the same numbers were learned together with the principle of their organization: the difference between 5 and 8 is 3, between 8 and 12 is 4, between 12 and 15 is 3, between 15 and 19 is 4, and so on. This principle was demonstrated by showing the numbers as follows:



Subjects learning the series in this fashion could grasp it very quickly and, indeed, could quickly learn a much longer series organized according to the same principle. One could easily construct such a series that would take a very long time to memorize by rote but which could be retained after a single exposure by someone familiar with the organizing principle. Katona, furthermore, found that his second group retained the series after an interim of a week.

When materials to be learned constitute a system it is possible to predict some of the materials from knowledge of others. Systematic learning occurs when principles are discovered which make it unnecessary to memorize detailed materials. The relevant whole

for these numbers is not the total series containing the individual numbers but is, rather, the principle governing the series. Systematic learning gives insight in that it provides principles (not always verbally formulated) from which specific materials can be derived. In learning to read there seems to be more insight provided by phonetic rules than by the look-and-say method. Learning to recognize the total appearance of a given word teaches nothing about recognizing other words. Each part is independent of all others. Learning is a process of memorization. When recurrent sound-letter matchings are learned we acquire a set of principles telling us how to pronounce indefinite numbers of new words; we learn the sound system of English writing. The fact that it is a very complicated and sometimes inconsistent system does not prevent it from being a system. Gestalt theory, then, would seem to favor the insightful phonetic method.

The use of whole words as teaching materials is as possible in phonetic training as in look-and-say recognition training. The best techniques for teaching phonetic generalizations (hereafter to be called the phonic methods) do work with whole words. Phonic training calls the attention of the student to words in which there are recurrent letters or groups of letters and correlated recurrent sounds. One might begin with a set of words all having the same initial letter in printed form and the same initial sound in spoken form: such a set as *mother*, *man*, and *milk*. From these words a general rule emerges: For the letter *m* make the sound heard initially in *mother*. A teacher using the phonic method will usually begin with the consonants since these have more consistent phonetic values than the vowels. With the vowels it is usual to teach the short forms first (as in *had*, *hen*, *hid*, *hod*, *hut*) since these are more common in English than are the long vowels (*hate*, *heed*, *hide*, *hoed*, *mute*). Still later come such contingent rules as the following: The sound /*k*/ is spelled *k* before *e* or *i* but it is spelled *c* before *a*, *o*, or *u* and *ck* after a short vowel. Finally there are some spellings for which no rule can be found and these are probably best taught last. All of the phonetic generalizations can be abstracted from words. They need not be taught by pronouncing individual letters. Certainly they will not be taught by reciting the alphabet. The names of letters (as opposed to their common phonetic values) must even-

tually be memorized since it is customary to spell by naming letters rather than by sounding them. But recitation of the alphabet is no part of good phonic preparation for reading.

If it is true that phonetic generalizations can be taught with whole words it is also true that pupils who are taught to recognize whole words can incidentally form phonetic generalizations and it is certain that most of them do so. This means that pupils do not dichotomize into two groups reading by entirely different methods. However, there are differences between the teacher who works by a phonic method and the teacher working by a look-and-say method. The phonics teacher will draw general rules out of words and she will explicitly state these rules from the first, encouraging her students to use them. The look-and-say teacher will provide materials from which general phonetic rules can be abstracted but, in the beginning, she will leave it to the student to find these rules. For the first year or two, at least, he must learn his phonetic generalizations incidentally, without explicit formulation by the teacher. Later on, the look-and-say teacher may institute some direct phonetic training. Oddly enough she has always been inclined to do so with backward children who need remedial reading help. The need for a phonetic attack on new words is generally recognized by educators of the look-and-say persuasion but, for one reason or another, they believe the necessary generalizations should be incidentally learned or, if directly taught, postponed until the second or third grade. What are the reasons for this belief?

Dolch and Bloomster have said: "It is true that the use of phonics means the use of generalizations, *that generalizations are best learned inductively*, and that sight words are the basis of inductive reasoning."¹ The italicized portion of this sentence is hardly a common sense observation. Why does the scientist write out his laws, the chef his recipes, the professional golfer his instructions for the novice if not to spare the rest of us inductive labor? We benefit from the experiences of our predecessors by reading the generalizations they formed. It may be that the Darwinian theory of evolution is best learned inductively—best in the sense of most unforgettably. But if it had to be learned that way most of us

¹ Italics my own.

would live without a theory of evolution. On the face of it a generalization is more rapidly and certainly learned when it is explicitly stated. In addition there are experimental results to show that incidental learning is slow and uncertain by comparison with directed learning. The educator who would claim that phonetic generalizations are better learned by incidental induction than by direct formulation with examples, assumes the burden of proof. His claim does not conform to popular belief nor has it been demonstrated in the laboratory. If you really want your pupil to learn a phonetic rule it seems sensible to tell him the rule.

Some educators think it best to teach phonetics directly, but argue that such training ought not to be used before the second grade. Until that time, it has been claimed, children have insufficient mental maturity to make use of abstract phonetic principles. Dolch and Bloomster found that first grade children taught by a look-and-say method failed to form phonetic generalizations which they could use in attacking new words. The authors conclude that a mental age of seven years, which usually means second grade standing, must be attained before a child can benefit from phonic training and that all such training ought to be postponed until he has reached that age. Quite obviously their results do not demonstrate that first grade children are unable to benefit from phonic training since the children of the study were not given *explicit* phonic training. First grade children know the rules of games that are fully as complicated as the rules involved in spelling. Furthermore, they are rather accomplished speakers of English, which means they have formed many concepts and learned complicated grammatical conventions. It seems unlikely that spelling rules are beyond them.

Empirical Evidence. We are not entirely dependent on theoretical argument and indirect evidence in deciding on the possibility of benefiting from direct phonic training in the first grade. In Scotland, children enter school at five years of age and begin the study of reading (by a phonic method) almost at once. In a study of the Committee on Reading of the Scottish Council for Research in Education children beginning the second year of school were given two American tests, the Metropolitan Reading Readiness Test and Primary I Battery, Form A, from the Metropolitan Achievement

Tests. The mean chronological age of the children was 6.3 years, approximately the age at which American children begin the first grade. The mean readiness score of the Scottish children was 90 which is 16.33 points above the norm for American children of the same age and I.Q. On the reading test their grade score was 7.5 which is at least a year above the norm for American children of the same age. It seems to be clear that children can gain from phonic training even before the first grade, and it would appear that a good way to build reading readiness is by instruction in reading.

There is also some experimental evidence on the relative merits of direct and incidental teaching for the development of phonetic knowledge. The basic design is always the same: One group is trained for a time using a phonic method and a more-or-less comparable group is trained for the same period by some non-phonetic method. Finally, both groups are given the same tests of reading achievement, and performances are compared. The studies vary greatly in the adequacy of their controls and in sample size. In summarizing this evidence Flesch ignored the significance levels of the differences found (as the authors themselves often do) and he missed several studies. Still, I find his general summary about right. Phonetic knowledge is more reliably acquired from direct tuition than by incidental induction from reading whole words.

Perhaps the three best studies are those of Agnew, Russell, and McDowell. Agnew compared all of the third graders in Raleigh, N.C., with a sample of 300 from Durham. In Raleigh a look-and-say method was used and in Durham a consistent, intensive phonetic instruction. The Durham children were reliably superior on oral reading, pronunciation, sounding letters, and pronouncing new words. Russell compared phonics-trained children with those trained by other methods at the end of the first year of reading instruction. His groups were equated for mental age. The classes that had received intensive phonetic training were significantly superior on eleven of twelve tests and particularly so in spelling, word recognition, and the sounding of letters. The subjects in the third research (McDowell) were fourth grade students from the Catholic schools in Pittsburgh. Five of the schools had, for three years, used intensive phonetic training while the other five had used a more general reading program in which phonetic training was one of

many kinds of instruction provided. The classes that had been given intensive phonetic training came out ahead on two tests—alphabetizing and spelling. Direct phonetic instruction produces superior skill in spelling, oral reading, sounding letters, and whatever aspects of reading call for phonetic knowledge. Incidental learning does not work as well.

We might have hoped for really conclusive evidence on the relative merits of the teaching methods from a comparison of reading achievement in the days when phonetic drill predominated with achievement today. However, there are no perfectly comparable data. Summarizing the best ten studies, Gray and Iverson decide that there has been no significant change in silent reading achievement in the past two or three decades. These authors add that average achievement in oral reading is not as high today as it was formerly because of radical changes in emphasis in teaching from oral reading to silent reading.

There remains one possible reason for avoiding intensive phonetic instruction and this reason is stronger than the others. We are asked to remember that the ability to sound out new words is not the only goal of reading instruction. In going after that objective with specific intensive training teachers may neglect reading speed, reading interest, and comprehension. There is some evidence in the experimental work that this can happen. Mosher and Newhall and Tate found look-and-say trained children slightly, but not significantly, superior to children trained by phonic methods on such tests as reading speed and silent comprehension. Agnew, in the study mentioned earlier, found the look-and-say trained children to be slightly more rapid readers. McDowell, in the study of Catholic schools, found the look-and-say classes superior on many tests but particularly so on paragraph comprehension, reading rate, and use of the Index. The sum of these results is that intensive phonetic instruction may take so much classroom time that other skills are slighted (e.g., use of the Index). Apparently, in recent decades, American education has been less concerned with phonetic knowledge than with other aspects of reading. Perhaps the loss in spelling and oral reading is more than compensated for by gains in comprehension and speed. However, on the basis of the experimental literature these gains appear to be slight, even doubtful. Gray and

Iverson, in their comparison of past and present reading achievement, are more certain that phonetic skills have declined than that other skills have improved.

Perhaps the most important test of the look-and-say methods is the volume and character of the reading done by those who learned to read in this way. If look-and-say has produced a generation that reads more books than past generations we will not worry if they also spell less expertly. Certainly more books are sold today than thirty years ago, about 68 per cent more than in 1929 according to the United States Department of Commerce. However, people today are spending more money on all of the mass media—on radio, television, and movies as well as books. Book sales today account for only 9 per cent of all recreational expenditures whereas in 1929 they represented 13.5 per cent, i.e., the rate of growth in book sales has been less than for competing means of communication. Does this change in book sales represent a debit or a credit for new methods of teaching reading?

There are many who will reject the volume of book sales as an index of interest in reading. How many routinely acquired books-of-the-month go unread? The volume of book reading, not book buying, is the relevant datum. Gray and Iverson report that comparative studies of sixth, seventh, and eighth grade pupils—thirty years ago and today—show no increase in reading. From the sixth to the twelfth grade the average number of books read in a month is now, as in 1925, about 1.5. Furthermore the chronological pattern of reading interest is much the same. It rises in the primary grades, reaches a peak in junior high school, and then falls off sharply in senior high school. The curve begins to rise a little earlier today than formerly but otherwise is about the same. As for today's adults, most of whom were taught by the new methods, according to DuShane, the American Institute of Public Opinion found that only 17 per cent were currently reading a book. The comparable figure in 1937 was 29 per cent. Perhaps the schools are doing well to maintain even this level since reading has today the fierce competition of television and movies. On the other hand, it must be acknowledged that children's books are much more attractive today than formerly and this change, rather than the change in methods of reading instruction, may be responsible for maintaining

children's interest. It is disappointing to find educators taking their stand on the claim that children today read as well as children have ever done. From the certain fact that this century has seen more discussion and research on reading than all prior centuries we should have hoped for something more than holding the line.

What To Do Now? The temperate, reasoned conclusion to this discussion, as to so many others, is that we won't know the answers until we have more and better research. The evidence certainly is not complete. We need longitudinal studies of matched groups of pupils trained by phonic and non-phonetic methods, studies comparing oral reading of new words, reading speed, interest, and comprehension. But the call for more research is a stale tune from the psychologist. There are some people, living in the real world, who have a reading class to teach today and another one tomorrow. What would you do, Mr. Psychologist, if you had to act today? The answer must lie in a combination of methods. I would begin phonic instruction in the first grade, not with recitation of the alphabet, but by extracting generalizations about the more consistent consonants from whole words in which the consonants appear. I would continue to stress meaning—combining the word with a picture when possible. I would do some flash card training with whole words, choosing for this purpose the common short words like *and*, *but*, *the*, etc. I would teach these words as total patterns because they are so common in English that it will greatly speed reading if they can be read in a glance. In addition these words, some of the oldest in the language, have many phonetic inconsistencies and so make poor material for first phonetic training and yet they cannot very well be postponed if children are to read stories. Therefore, I would teach them by the look-and-say method. In my phonic instruction I would use only those words that are familiar to the child in spoken form so that he might have the thrill of recognizing and understanding his first halting pronunciations. I would rely on the satisfactions involved in such recognition to make reading interesting. This procedure that I have espoused involves more phonics at an earlier age than is now customary in American education. In taking such a stand I am sorry to be allied with unreasonable parents and enemies of progressive education. If we have all arrived at somewhat similar

views on this issue I think we got there by different routes and our ways are likely to diverge again.

Probably some parents are angry because they feel guilty about not doing much reading themselves. For some reason the act of reading is held to be virtuous in its own right. If that statement needs documentation it can be found in Berelson's article on "What Missing the Newspaper Means." There is supposed to be a virtue in reading that does not accrue to listening or viewing and the printed word enjoys a kind of approval that is not given to radio or television. This feeling is very common among college teachers who would feel personally discredited by the presence of a television set in their homes. Probably it is nonsense to enthrone one medium (the book) and despise another (the television). There is enough trash to go around. I don't think it can be shown that quality content is limited to one rather than another. Certainly there is nothing in the nature of the medium that prevents radio or television from being highbrow. However, whether sensible or not, reading is held to be a virtuous act (more or less regardless of the content) while viewing and listening are relaxing but a waste of time. The parent who is guiltily conscious of the fact that he does rather more viewing than reading may be distressed to see Johnny going to hell in the same way. His guilty anger can be handily displaced to the teacher and the look-and-say method that prevents Johnny from being a reader. The teacher has a strong point when he argues that children are not going to be eager to learn to read if they live among people who seem to get along very nicely without opening a book. If the Johnny of such parents has trouble with reading surely the parents ought to inquire into the extent of their own responsibility. Whether they choose to regret his backwardness and feel guilty about it is a matter between them and their culture.

The teacher has reason on his side again when he refuses to be overwhelmed by a parent's memories of his own great interest and rapid progress in reading as taught by the phonic method. There is more than one variable here. Such parents may be of superior I.Q. and have had higher mental ages than their classmates or the whole reminiscence may be rosy tinted by nostalgia. Neither should the teacher change his methods because some parent—often a college professor—tells him of his great success in teaching Johnny to read

at home by using a phonic method. I have heard wonderful stories of this kind with children learning to read at four years of age, chronological age that is. But the child of a college professor probably has a high I.Q. and so may have a mental age of five or six when he is four years old. Such a child might be particularly apt at learning phonics or most anything else. The methods of the public schools cannot be geared to them. There are great individual differences in the abilities of pupils in the primary grades. Most of them will learn to read by present methods. Most of them will make phonetic generalizations for themselves whether or not the teacher points them out. I think, however, that more of them would learn to read better and sooner with more explicit phonic instruction.

SUMMARY

To learn a written name for each referent category is a big job and writing systems all provide some kind of short cut to this knowledge. The earliest systems took advantage of the psychological economy in representation. The symbol manifested some criterial attributes of the referent and so suggested the referent. For various reasons this economy had a quite restricted usefulness, probably more restricted as societies grew in size and complexity. The written form of a language provides names for the same referents as does the spoken form, and the spoken names are generally learned first. This fact makes another economy available to a writing system. The phonetic writing, whether syllabic or alphabetic, translates recurrent speech elements into written characters and combines the characters into names as the sounds are combined in speech. When one learns such an alphabet or syllabary he ought to be able to read, write, and spell all the names that are familiar to him in spoken form. It is an irony of history that this economy, which made the invention of the alphabet so important to mankind, has been partially lost to such languages as English and French. English orthography is today a very inconsistent phonetic system. This fact has suggested to many American educators that literacy in English ought to be taught without explicit reference to the phonetic values of the letters. However, the evidence indicates that teachers do better to call attention to the phonetic system that exists, even though it is exasperatingly irregular.

Logic, Thinking, and Teaching*

Logic is one of the ways of studying language usage. This selection, then, is appropriate to the general consideration of communication. Its content, the reader will discover, is relevant also to the previous chapter on cognitive learning and may, in some respects, be considered an attack on the research in that section.

An article which criticizes an "over-psychologizing" of teaching, especially the way *thinking* is taught, may be especially helpful in establishing the limits of psychological knowledge and method and in enabling one to choose more wisely and cautiously the principles and techniques which may prove most useful in educational research and in the improvement of educational practice. Smith's attack on experimental psychologists such as Thorndike is a stronger reiteration of the opinion stated by Melton (pp. 6-7). Experimentalists have frequently confined their investigation to observable physical behaviors, especially in animals of limited symbolic power, and have ignored the verbal behavior of humans. Crowder maintains that this is the great weakness in Skinner's position (pp. 186-187). This, of course, imposes great restrictions on what the educational psychologist can borrow from experimental psychology. For, as the author states, "the plain fact is that without language, nothing can be taught or learned about the past, nor about things removed from immediate observation." Most teaching involves verbal discourse between the teacher and the student. According to Smith, it is a game of thinking in which the rules (that is, logic) are never taught—often because teachers are unaware of them.

* Reprinted and abridged with the permission of the author and publisher from the article of the same title, *Education Theory*, 7 (1957), 225-233.

The article is a controversial one; the student can analyze it by using these questions: (1) What distinction does the author make between logic and the psychological study of thinking? (2) Which (logic or psychology) most likely describes the laws of thought as compared to the laws for thought? (3) Why does Smith make a severe attack on Dewey's model for problem-solving? In what way would this attack also include the views of Bruner (pp. 254-270)? (4) What is the relationship, as it appears in this article, of logic to scientific method? (5) How convincing is Smith when he argues that education students should be required to have a logic course? (6) How does his view of experimental psychology agree with Crowder's (p. 186)?

My purpose is to explore the proposition that logic is relevant to thinking and teaching, and that preparation of the teacher should include the study of what I shall call educational logic.

LOGIC AND RIGOROUS THINKING

Had the normative significance of Dewey's theory of inquiry been understood in pedagogical circles, the course of educational development might have been different. For one thing, we might have had a logical basis for developing in students the capacity to direct and control their own thinking. As it was, the slogan that we teach "how to think" and not "what to think" was largely empty. For the pattern of inquiry which Dewey had earlier called the Complete Act of Thought was taken to be a description of how thinking and learning in fact occur, and not as a set of norms for judging inquiry. The stripped-down model of thinking presently found in educational psychology texts and other pedagogical literature consists of the following phases: feeling uncertain, locating barriers to action, getting hunches and trying them out in imagination, acting upon the promising one, and deciding whether the action ended with the desired result. Now these are psychological. They describe what we do when as hungry cats we try to escape from our box. They may be fairly good descriptions of what we do,

but the crucial point is left out—the basis for appraising results. Blind imitation of past performance is one thing, judging and improving it, another.

There is another important point: from the standpoint of psychology, these five steps do not necessarily involve the use of language; for behavioristic theory assumes that observable actions and not talking are the true index of psychological events. What men say is one thing and what they do, quite another. The psychologist trusts in the doing and rules out language as significant behavior. This mistrust of language is due partly to the fact that experimental psychology developed largely from the study of lower animals and partly to the discredited status of language as a result of its use in introspection which was repudiated. In any event, the mistrust of language by psychologists infected educational thought in spite of the obvious fact that both teaching and human learning are practically impossible without it.

If we look in educational literature for criteria by which to evaluate the steps we take to get out of the box, we find that we are admonished to be alert, careful, open-minded, and whole-hearted. This is not very helpful. In the first place, these terms are elusive and hard to pin down. How can we tell when we are sufficiently alert, let alone, when someone else is alert? If A says he is alert and B says no, A is not alert, by what criterion can we decide whose claim to accept? The same vagueness appears in the advice to be careful, open-minded, and whole-hearted. In the second place, these exhortations refer to ways of acting rather than to statements in which ideas can be formulated and tested. We can say that an individual is alert, careful, and so on. But it is a bit odd to say of a statement that it is alert, careful, whole-hearted, or open-minded.

Yet, since much of what we do—especially in matters of reasoning—is done with language, these vague psychological criteria are practically useless. They are useless in dealing with verbal formulations by which the logical soundness of thinking may be tested. Since they refer to ways of acting, these criteria are part and parcel of the psychological version of problem-solving, where results are tested in terms of success, and where success means to find a satisfying way out of a difficulty. But a mere state of psychological satisfaction offers no guarantee that we are not mistaken, nor that the way we

solved the present problem might not lead to unsatisfactory consequences the next time we try it. Thus, the surrender of intellectual standards to psychology has all but robbed education of a theory of rigorous thinking.

By virtue of language and logic, thinking takes on a dimension in man different from that of the cat in the box. Man is a language-using creature; this fact opens to him spheres of experience not given to other creatures. A point so obvious might go without saying were it not for the fact that the epithets "merely verbal" and "verbalism" have dulled the edge of understanding where language and logic are involved in teaching and learning. The plain fact is that without language, nothing can be taught or learned about the past, nor about things removed from immediate observation. The laws of science can be learned only through language and retained in symbolic form alone. Without language the scientific method could not progress beyond the scramblings of the cat in the box. An adequate theory of control over our thinking will acknowledge the central role of linguistic behavior.

But unless a theory of disciplined thinking is based on logic, it is apt to emphasize language in its literary and sociable uses, and to neglect it as an instrument for directing the exploration of the environment and as a vehicle of knowledge. When this happens, the existential and logical monitors of our judgment are bypassed. Instead, we are exhorted to control our emotions—to prevent ourselves from being swayed unduly by prejudice and from losing our heads in anger, fear, or envy. But how can one tell whether or not he is controlling his feelings and prejudices, or whether someone else is doing so? Unless we are to fall back upon such vague and shifting psychological criteria as satisfaction and success, we must appeal to the norms of inductive and deductive logic. We know that one's feelings and prejudices do not interfere unduly with his thinking if he takes such precautions as these: uses a fair sample, recognizes what is true by definition, avoids fallacious reasoning, reduces ideas to observation sentences when an empirical test is to be used, and controls relevant conditions. The point of this analysis is that logical and not psychological criteria are used to judge the efficiency of such operations as these. Therefore the individual is disciplined in

his thinking only to the extent that he has taken on the logical and linguistic criteria by which the intellectual work of man is brought under control.

Now I want to explore the ways in which the thinking of an individual can be said to satisfy logical criteria. His thinking meets these criteria, we would say, when the results of his verbal behavior—statements, arguments, descriptions, and so forth—correspond to the rules. But the matter is not so simple, for there are different ways in which verbal behavior is found to abide by these rules. Consider the case of a child learning his mother tongue. As he does so, the child's sentence structure will conform to that of the adults. The child is not aware of such things as linguistic rules. In the same way, the child's reasoning may be said to conform to the rules of logic. His sentences may express valid arguments. This is rarely the case at an early age, but as the child progresses through the elementary school, he begins, though unconsciously and irregularly, to take on the forms of valid reasoning. In conforming to the rules, the child's reasoning is valid even though the idea never occurs to him that he should reason validly. He cannot verbalize the rules to which his reasoning corresponds. He has no words for talking about the rules of thinking or about mistakes in thinking. Hence his reasoning has not reached the threshold of conscious control. Now in cases of this sort we can say that the extent to which rules are satisfied at all, they are satisfied only by unconscious accommodation of behavior.

Let us now suppose that one who has learned to speak his native language goes to school and begins to study grammar. He is taught to classify words and to see relations among them in sentences. This gives him a set of words and rules for talking about language. He can then examine his own discourse and that of others to find and correct points where it fails to satisfy grammatical rules. But just as he can learn to use rules of grammar to guide his use of language, so can the individual learn to govern his own thinking and experimental investigations by using the rules of reasoning and inquiry. He thereby acquires greater power over his own intellectual activities and clearer insight into those of others. For now he can rise above, survey, and assess them in the same sense that a referee rises above the game to judge the plays. This ability to perform at more than one level of the intellectual game, both to play it and to stand

off and judge it, frees our thinking from unconscious habits conditioned by chance circumstances. Now when the individual examines his thought and brings it into line with criteria of effective thinking, we can say that this is a case of deliberate control, and is in direct contrast to unconscious accommodation.

Now I wish to make two qualifications. First, I do not wish to say that the individual is to make conscious use of logical rules every moment of his thinking. Constant mindfulness of these rules would be a serious impediment. It would have the same effect upon the thoughtful behavior of the individual as awareness of his feet was supposed to have had upon the march of the centipede. No one can act effectively if he is required at the same time both to perform and to think about whether or not he is performing correctly. But there are occasions when an activity can be improved by paying attention to its performance. Thus, one who is able to make deliberate use of rules when judgments about his own thinking or that of someone else are called for, is more apt to catch defects than one who has no command of logic. The disposition to use the rules when they count rather than the constant use of them distinguishes the discerning critic from the crass one.

The second qualification is about the rules themselves. It is usually thought that when one speaks of rules, he is talking about statements such as those in official rules books. But of course this need not be the case. Knowing rules does not mean knowing a particular set of words. The controversy over the teaching of formal grammar was not about the question of whether or not to teach grammatical rules. It was over the question of which rules to teach as well as the context in which the rules were to be taught, whether to teach them in relation to writing and speaking or as formally ordered elements in the official rule book.

When does one know the rules of a game, say, the game of checkers? Suppose he says, "No, you can't make that move" when someone moves a man backward, and he says this on any occasion when that move is made. Should we say that he knows one of the rules of the game, even though he cannot give a sophisticated statement of it? I think we should claim that he does know the rule. So it is with the rules of logic. To know that affirming the consequent is invalid is to recognize such affirmation when it occurs and to recognize thereby

that the truth or falsity of the conclusion is still up in the air. The particular verbal form in which the rule is put is of no consequence in playing the game or in refereeing it.

The analogy between the game of checkers and thinking breaks down at least in one significant respect. A player in checkers is always called for infraction of the rules; learning to play the game entails learning the rules. But with thinking it is different. A player in this game, except among professionals, can take all sorts of liberties without anyone calling him for infraction of the rules and without the player himself even knowing that he is breaking them. In some cases, however, he will pick up certain logical rules in an *ad hoc* sense. Suppose a beginning high school student is given the following argument: If it rains, the streets are wet. The streets are wet. Therefore it rained. He will tell you quickly that the conclusion does not necessarily follow because the streets may be wet for some other reason. Perhaps the street sprinkler has come along. But when the content is unfamiliar and the argument complex, the student will seldom recognize the fallacy. He may fail to recognize the fallacy as such, if his reasoning is still at the level of concrete relations. Hence he could not go beyond cases of particular content. Nor could he recognize the fallacy in any general sense. Hence if the material relations in an argument go beyond his concrete knowledge, the student who has only an *ad hoc* command of the rules cannot detect logical mistakes.

To recognize fallacies in a generalized sense is to see them in abstraction from any particular content. The individual is able to lay his finger on any instance of a particular fallacy and to tell what is wrong, even though he is not familiar with the subject matter. He can thereby use logical rules to monitor his own thinking as well as that of others. It does not require a formal study of logic and scientific method to give the individual such generalized control over his thinking and judgment. To be good at the game of thinking, one does not need to be a professional. At the professional level we find logicians and critics of logical and scientific systems. Here the individual can examine relations among rules, study them as a system, and propose changes in them.

To sum up the levels at which rules may be known, the first is the level of accommodation. At this level, the individual knows how to

play the game of thinking, though he does not explicitly know the rules. At the next, the *ad hoc* level, the individual knows the rules only in a special context. At the third, or general level, we now add the ability to deal with the rules in their general applicability. Finally, at the professional level, the individual goes beyond the general grasp of the rules to a command of them as a formal system.

I believe that at best the rules of the game of thinking are now being learned in an *ad hoc* sense. Both in school and out, the individual simply picks up rules of reasoning from the material relations in the content he studies. I suspect that many persons play the game of thinking without even knowing that there are rules. The teacher plays it too, and often on the same level, checking the students' thinking by the textbook, by what the teacher has himself been taught to be the correct answer, or by his own common sense notions of clear and accurate thinking.

LOGIC AND TEACHING

To continue a little longer in the metaphor, in the game of thinking the teacher is player, coach, and often referee. As a player, he engages students in thinking by asking questions and responding to their answers, by receiving questions and giving answers, and by many other devices and activities. In each of these there is a sort of give and take between teacher and students. But, having little knowledge of logic and being preoccupied with getting the student to understand facts and ideas, the teacher usually overlooks the logic of both his subject and of the class discussion. For instance, a history teacher discusses with his students the imperialism of a nation. He goes into the question of the extent and cause of the imperialism. But the concept of imperialism is not itself explicated, so that the students have varied notions of what is being talked about. The whole discussion is based on a vague and ambiguous term and thus thinking and learning are short-changed.

Now the teacher moves from the role of player to that of coach when he turns to the task of helping students work out a definition of imperialism. To handle this task, the teacher needs criteria by which to decide the adequacy of the definition worked out by himself and the students. As the teacher and students together analyze

the concept of imperialism and give it the form of a definition, the teacher will help students from time to time to see what it means to define a term and to understand the kinds of rules by which the adequacy of a definition may be decided. He will show them, as appropriate occasions arise, that a definition lays down criteria for the use of a word, and that the definition we decide to give a word, or the usage we select, is related to the purpose we have in mind. He will show them that sometimes we define words by assigning whatever is named by the word to a class and then distinguishing it from other members of the class. On other occasions the teacher will show how to define words by pointing to instances, and in still other cases by reference to the operations we perform.

To reflect upon the work of the teacher is to see that there are many occasions when he could readily teach procedures of analysis and logical appraisal. He could teach how to distinguish between logical validity and empirical truth, how to tell when an argument is valid and when a proposition is true. Through his instruction students could learn how to identify assumptions, how to tell when conditions are logically necessary and sufficient, how to identify hypotheses and to tell whether or not they are confirmed by particular instances, and many other procedures too numerous to mention. In order to teach for effective thinking at this higher level of operation, the teacher must possess a working knowledge of logic in an amount far in excess of that picked up through his own incidental learning. This claim that the teacher should be trained in logic rests upon two premises: first, it is important to develop the student's ability to think critically; second, in order to develop this ability the student must be given experience in controlling his own thinking under the guidance of the teacher.

Logic is not only necessary in teaching the student how to control his thinking but it is also an inherent part of teaching. Two lines of reasoning suggest this conclusion. In the first place, instruction in knowledge starts with an intent to arrive somewhere, to reach a conclusion. It is not a form of sociable conversation meandering wherever fancy leads. Rather it is studied discourse which aims not only at conclusions but also at showing the steps and reasons leading to the conclusions. This kind of discourse is logical. In trying to teach students to understand a conclusion and how it is reached, a teacher

does not lead them down the blind alleys, detours, and mistaken paths which he himself may have taken, nor does he expose them to the false leads taken by the race. He strips off these mistaken moves, and starting at some take-off point, he tries to show by a chain of ideas that the conclusion is warranted.

In the second place, observation bears out the fact that instruction in knowledge involves logic. Even cursory observation of teaching will show that the teacher performs certain logical operations. He defines, interprets, explains, justifies, proves, evaluates. Now each of these activities is an operation done with words, sentences, and statements, and which we cannot perform without these linguistic instruments. Thus when we explain an event in science, we show it to be a special case of a law. In proving a proposition in mathematics, we show it to be a conclusion from a set of premises. When justifying our action, we give reasons to show that it is a wise step to take. In defining a word, we state the rules for using it. Now all of these operations fall in the domain of logic.

These operations are to be sharply distinguished from psychological processes. To talk about the operations of defining, proving, explaining, and so on, is not to talk about psychological events. Psychologists talk about certain processes: perceiving, emoting, conceiving, inferring, judging. Now it may well be that these processes occur when we prove a proposition or justify a course of action. But I do not think that teachers work with such processes even though they may be going on. Rather the teacher works with words and statements, their meanings and relations. He may decide that certain psychological events are taking place at a given moment, and so modify what he is doing. Still he is dealing with signs and symbols and performing the logical operations entailed by verbal instruction. So the changes he does make in his activities are largely changes in verbal behavior.

Some critics might object that my analysis presupposes an old-fashioned way of teaching and would then go on to make the point that nowadays no good teacher uses the methods of talking and telling. Instead of such verbal methods, these critics would say, the teacher engages students in problem-solving activities where they learn by planning and working out things together. It would then be said that the methods of involving and directing students in these

activities require no special knowledge of logic on the part of the teacher. What he needs instead is suitable knowledge of human relations and skill in the techniques of group organization and control.

But it would be easy to show that the usual practice of teaching, especially in high school and college, is neither group work nor problem solving. Then, too, it can be shown that the assertion "a good teacher uses such and such methods" is logically circular because the meaning of "good teacher" may be interpreted to include "uses such and such methods." It is like saying all black cats are black. However, since these replies are not likely to convince the critics, I shall forego the pleasure of playing around with them.

Yet it should be noted that this objection to the claim that teachers need command of logic does not touch upon the point that the control of thinking involves deliberate use of logical rules. Such objection bears only on the claim that the acts of teaching are themselves logical operations. But, suppose it were the case that all teachers use problem-solving and group work. What would the teacher do in the classroom? If one insists that all learning is to result from problem-solving, or from instruction by one's peers through group work, what is the teacher's role? Is he no more than a mechanic who pushes psychological buttons to activate students and to keep them from getting out of hand? Of course not. The teacher who uses problem-solving and group work is inescapably involved in all the talking and telling that normally goes on in teaching. He must work with individuals and groups helping them to clarify meanings, to analyze and evaluate reasons and arguments, hypotheses and plans of action. The teacher helps students in testing the truth of statements about matters of act, and in dealing with many other linguistic and logical operations. Even in problem-solving and group work the teacher cannot do these things without talking, and he will be intelligent about his talking only to the extent that he exercises deliberate control over it.

Verbal Context and the Recall of Meaningful Material *

Linguists have told us that any language is based on patterns of speech and that every language favors some patterns to the exclusion of many others. From early childhood we hear and learn these patterns, even before we fully "understand" them in the syntactical and grammatical sense. Yet we develop an un verbalized feeling for these patterns and recognize adherence to or departure from them. Such findings by linguists have already altered the teaching of language in our schools, where the emphasis sometimes has shifted from visual to aural training.

The following article illustrates how experimental psychologists study language by using statistical analysis. The authors define verbal context in terms of "dependent probabilities." In general usage there is less likelihood (or probability) that word *Z* will occur after word *W* than word *Y*. The investigation shows how a statistical English may be constructed by using a "musical chairs" procedure. The lists compiled by the authors ranged from apparent nonsense to meaningful sentences (see pp. 377-380). In this particular study the experimenters wanted to know how these various levels of meaningfulness influenced recall. Their results may come as a surprise. Questions that are important here are: (1) What did the authors discover about the recall of these materials? (2) How do they explain these results? (3) What concept would they substitute for meaningfulness? (4) In view of this experiment how would you define rote learning? (5) As a teacher you will want to employ methods which assure maximum recall for your stu-

* Reprinted with permission of the senior author and the publisher from article of same title, *The American Journal of Psychology*, 63, 1950, 176-185.

in terms of dependent probabilities.³ The probability that event *C* will occur is not the same after *A* as it is after *B*. The statistical dependencies between successive units form the basis for a study of verbal context.

To illustrate the operation of conditional probabilities in our verbal behavior, consider the set of all possible sequences 10 letters long. We could construct a table listing them. The first row of the table would be the pattern *aaaaaaaaaa*, 10 consecutive *a*'s. The second would be *aaaaaaaaab*, then *aaaaaaaaaba*, *aaaaaaaaabb*, *aaaaaaaaabaa*, etc., until all possible arrangements of letters, spaces, commas, periods, hyphens, quotes, colons, numbers, etc., were exhausted. Altogether there would be about 50 different symbols, and the table would contain 50^{10} , or about 100,000,000,000,000,000 different patterns. Then we would examine some English writing and try to determine the relative frequencies of occurrence of the patterns. Only a small fraction of the 50^{10} alternatives actually occur in English. The table would show strong dependencies. For example, the letter *q* is always followed in English by the letter *u*, and so all those entries in the table that contained a *q* followed by anything but *u* would not occur in English. It is not possible to predict the relative frequency of *qe*, for instance, by multiplying the relative frequencies of *q* and of *e*.

If such a table existed, along with the relative frequencies of occurrence, it would be possible to construct sequences of letters that reflected the statistical dependencies of English verbal patterns. We can imagine similar tables constructed for shorter or longer sequences of letters. A table for all patterns of 2 symbols would represent the relative frequencies of pairs; for 3 symbols, triads, etc. The longer the sequence, the more information the table contains about the pattern of dependencies in our molar verbalizations.

To illustrate the use of such information we shall borrow a device used by Shannon. Suppose we have no knowledge at all of the relative frequencies of occurrence, but only a list of the 50 different symbols. Then, for all we know, any sequence of symbols might be permissible. If we tried to construct a message in the language, the best we could do would be to draw at random from the 50 different symbols. We have no reason to think that one sequence of symbols is more likely than another. Proceeding in ignorance to construct a message, we might produce something like this: *cplp'rw(p'':k!)"ntegznqO?i6-vlaur :8h*, etc.

Now suppose that we have a reliable tabulation of the relative frequencies of "patterns" of one symbol. We know, therefore, that *e* and

³ G. A. Miller and F. C. Frick, Statistical behavioristics and sequences of responses, *Psychol. Rev.*, 56, 1949, 311-324.

the space between words are more likely to occur than are ? and z. With this information we can increase the chance of constructing a meaningful message, although our chances are still very small. If we draw successive symbols according to their relative frequencies of occurrence in English, we might produce something like this: *wli hnrooye lricocnri mae c zg 2eaya*, etc.

The next step is to imagine that we have a tabulation of relative frequencies of occurrence of patterns of two symbols. Now it is possible to improve the statistical approximation to English by drawing in the following way. Begin by drawing any likely pair. Suppose the pair is *au*. Now look at all the pairs starting with *u* and draw from them according to their relative frequencies of occurrence. Suppose the result is *ud*. Now look at all pairs starting with *d* and draw one of those, and so proceed to build up the message. Notice that each draw depends upon the preceding draw—the preceding draw determines from which set the present draw is to be made. Drawing in this way reflects the conditional probabilities of successive symbols. A message constructed in this way might read *aud ren stiofivo omerk. thed thes bllale*, etc.

If we have a tabulation of sequences of three letters, we can construct a message that reflects the conditional probabilities of English triads. First we draw a likely triplet, say *ann*, then draw next from the triplets starting with *nn* and obtain *nna*, then from the triplets beginning *na*, etc. The preceding two symbols determine from which set the next triplet is drawn. In this way a message might be produced that would read: *annation ef to the acticas. Oth rested*, etc.

With a tabulation of sequences of four letters we might produce: *influst intradio be decay, the condive*, etc. By tabulating the relative frequencies of longer sequences and drawing successive items so as to reflect these frequencies, we can construct messages that reflect the statistical dependencies of English as extensively as we please.

For convenience, we shall refer to these different ways of constructing a statistical English as orders of approximation, and shall number them from 0 to *n*. At the zero order we have no knowledge of relative frequencies, at the first order we know the relative frequencies of individual symbols, at the second order we know the relative frequencies of pairs, at the *n*th order we know the relative frequencies of *n*.

Consider this statistical English now in terms of verbal context. With a zero-order approximation to English there are no contextual influences whatsoever on the choice of successive symbols. At the *n*th-order approximation, however, each symbol is selected in the context of the preceding *n* - 1 symbols. As the order of approximation is increased, the amount of context for each symbol is increased, and the contextual constraints (dependent probabilities) have a chance to

operate. As the order of approximation is increased, the messages we can construct become more and more familiar, reasonable, meaningful. The more we permit contextual restraints to operate, the better are our chances of producing a message that might actually occur in English.

We have, therefore, a scale for what can be loosely called 'meaningfulness.' At one end are the random jumbles of symbols we customarily call nonsense, and at the other end are patterns of symbols that could easily appear in our daily discourse. Equipped with this quantitative estimate of 'the degree of nonsense' or 'amount of contextual constraint,' we can proceed to study certain psychological problems that have been phrased in terms of meaningfulness.

An Experimental Illustration. Briefly stated, the problem to which this concept of verbal context has been applied is, How well can people remember sequences of symbols that have various degrees of contextual constraint in their composition? The experimental literature contains considerable evidence to support the reasonable belief that nonsense is harder to remember than sense. This evidence has suffered, however, from a necessarily subjective interpretation of what was sensible.

In the present experiment, the learning materials were constructed at several orders of approximation to English. These materials were presented to Ss whose recall scores were then plotted as a function of the order of approximation.⁴

Learning materials. In the preceding examples we have used patterns of letters to illustrate the effects of contextual constraints. There is, of course, no necessity to limit the argument to letters. It is possible to use words or even sentences as the component elements that are arranged according to the statistical structure of English. In the present experimental illustration the materials were constructed with words as the units of analysis.

In theory, the construction of materials to incorporate the statistical structure of English over sequences of several words requires a tabulation of the relative frequencies of such sequences. Such a tabulation would be exceedingly long and tedious to compile. An alternative method of construction is available, however, which makes the procedure practicable. Instead of drawing each successive word from a

⁴ The experiment was carried out by the junior author and is presented in detail in her honor's thesis, *Investigations into the Structure of Verbal Context*, 1949. The thesis is on file in the Library of the Psychological Laboratory, Memorial Hall, Harvard University.

different statistical distribution indicated by the preceding words, we draw the word from a different person who has seen the preceding words.

At the second order, for example, a common word, such as *he*, *it* or *the*, is presented to a person who is instructed to use the word in a sentence. The word he uses directly after the one given him is then noted and later presented to another person who has not heard the sentence given by the first person, and he, in turn, is asked to use that word in a sentence. The word he uses directly after the one given him is then noted and later given to yet another person. This procedure is repeated until the total sequence of words is of the desired length. Each successive pair of words could go together in a sentence. Each word is determined in the context of only one preceding word.

For higher orders of approximation the person would see a sequence of words and would use the sequence in a sentence. Then the word he used directly after the sequence would be added, the first word of the sequence would be dropped, and the new (but overlapping) sequence would be presented to the next person. By this procedure we constructed sequences of words at the second, third, fourth, fifth and seventh orders of approximation.

For the first order approximation to English a scrambling of the words in the higher orders was used. By drawing words at random from the contextually determined lists, we obtained as good an approximation to the relative frequencies of individual words in English as these higher order lists provided. The alternative method of selecting words at random from a newspaper might have given a sample quite different in difficulty (familiarity).

A zero order approximation to English could be obtained by drawing at random from a dictionary. Most dictionaries contain too many rare words, however, so we drew from the 30,000 commonest words listed by Thorndike and Lorge.⁵ This source had the additional advantage that it listed separately all forms of the word, whereas the dictionary lists only the lexical units. Words drawn at random from this list of 30,000 words are selected independently and without any constraints due to adjacent words or the relative frequencies of appearance of the words in English.

A final set of words was taken directly from current fiction or biography. These lists represent a full contextual determination.

By these devices we constructed sequences of words with eight different degrees of contextual constraint. In the following discussion we shall refer to these lists as 0, 1, 2, 3, 4, 5, 7 and text-orders of approximation. At each order four lists of different length—10, 20, 30 and 50 words—were constructed. Thus the experimental design called

⁵ E. L. Thorndike and I. Lorge, *The Teacher's Wordbook of 30,000 Words*, 1944.

creases as the length of the list is increased. Inspection of Fig. 1 leads to a reasonable suspicion that the two variables, length and order of approximation, interact. With the short, 10-word lists there is little to be gained from contextual bonds extending over more than two words. With the 20-word lists the Ss remembered as well at the third order of approximation as they did for the textual material. With the 50-word lists, however, only orders 5 and 7 are comparable to the textual material in terms of percentage recalled. It would seem, therefore, that the longer the passage the greater is the usefulness of contextual associations extending over long sequences of items.

By a strict interpretation of the word "nonsense," one is forced to conclude that all orders of approximation less than the full text are nonsense. Consider an example from Order 5:

house to ask for is to earn our living by working towards a goal for his team in old New-York was a wonderful place wasn't it even pleasant to talk about and laugh hard when he tells lies he should not tell me the reason why you are is evident.

The experimental results show that this kind of gibberish is as easily recalled as a passage lifted from a novel. Thus there are kinds of nonsense that are as easy to recall as are meaningful passages. The significant distinction is not to be drawn between meaning and nonsense, but between materials that utilize previous learning and permit positive transfer and materials that do not. If the nonsense preserves the short range associations of the English language that are so familiar to us, the nonsense is easy to learn.

The experiment shows, therefore, that the problem of meaning vs. nonsense in verbal learning need not be approached in terms of a qualitative dichotomy, but can be studied as a functional relation between quantitative variables. The results indicate that meaningful material is easy to learn, not because it is meaningful *per se*, but because it preserves the short range associations that are familiar to the Ss. Nonsense materials that retain these short range associations are also easy to learn. By shifting the problem from "meaning" to "degree of contextual constraint" the whole area is reopened to experimental investigation.

Psychologists familiar with the problems of verbal learning will recognize the usefulness of the kind of material employed in this

illustrative experiment. For example, is retroactive inhibition affected by interpolating different orders of approximation to English between the original learning and the recall? What is the effect of using original and interpolated materials of the same or of different orders of approximation to English? Do the higher approximations to English show the same differences between recall after sleep and recall after waking activity that the lower approximations show? Is it possible to show a continuum from the short-term reminiscence that can be demonstrated with syllables to the long-term reminiscence that can be shown with poetry? How does the span of immediate memory vary with the order of approximation? Is the superiority of distributed over massed practice a function of the order of approximation of the materials to the statistical structure of English? Can differences in learning and recalling different orders of approximation be demonstrated as a function of age?

The operational analysis of meaningfulness makes it possible to ask such questions and to see how one would proceed to answer them. The problem now is to collect the experimental data.

Summary

A quantitative definition for verbal context is given in terms of dependent probabilities. The definition is used to construct lists of words with varying degrees of contextual determination. When short range contextual dependencies are preserved in nonsense material, the nonsense is as readily recalled as is meaningful material. From this result it is argued that contextual dependencies extending over five or six words permit positive transfer, and that it is these familiar dependencies, rather than the meaning *per se*, that facilitate learning.

Appendix

LISTS USED IN RECALL EXPERIMENT

0-order approximation

- 10: byway consequence handsomely financier bent flux cavalry
 swiftness weather-beaten extent
- 20: betwixt trumpeter pebbly complication vigorous tipples careen

- obscure attractive consequence expedition pane unpunished
 prominence chest sweetly basin awoke photographer ungrateful
 30: crane therewith egg journey applied crept burnish pound
 precipice king eat sinister descend cab Idaho baron alcohol in-
 equality Illinois benefactor forget lethargy fluted watchtower
 attendance obeisance cordiality dip prolong bedraggle
 50: hammer neatly unearned ill-treat earldom turkey that valve
 outpost broaden isolation solemnity lurk far-sighted Britain
 latitude task pub excessively chafe competence doubtless tether
 backward query exponent prose resourcefulness intermittently
 auburn Hawaii unhabit topsail nestle raisin liner communist
 Canada debauchery engulf appraise mirage loop referendum
 dowager absolutely towering aqueous lunatic problem

1-order approximation

- 10: abilities with that beside I for waltz you the sewing
 20: tea realizing most so the together home and for were wanted to
 concert I posted he her it the walked
 30: house reins women brought screaming especially much was said
 cake love that school to a they in is the home think with are
 his before want square of the wants
 50: especially is eat objections are covering seemed the family I
 that substance dinner raining into black the see for will pas-
 sionately and so I after is window to down hold to boy appear-
 ance think with again room the beat go in there beside some is
 was after women dinner chorus

2-order approximation

- 10: was he went to the newspaper is in deep and
 20: sun was nice dormitory is I like chocolate cake but I think that
 book is he wants to school there
 30: the book was going home life is on the wall of you are ready
 to the waltz is I know much ado about it was a dog when it was
 50: you come through my appetite is that game since he lives in
 school is jumping and wanted help call him well and substance
 was a piano is a mistake on this is warm glow in and girl went
 to write four turtledoves in my book is fine appearance of the

3-order approximation

- 10: tall and thin boy is a biped is the beat
 20: family was large dark animal came roaring down the middle of
 my friends love books passionately every kiss is fine
 30: happened to see Europe again is that trip to the end is coming

- here tomorrow after the packages arrived yesterday brought good cheer at Christmas it is raining outside as
- 50: came from the beginning and end this here is the top spins in a house by the library is full of happiness and love is very nice of her that fell from the window she went home from work to pass the cigarettes down to earth he picked an apple

4-order approximation

- 10: saw the football game will end at midnight on January
- 20: went to the movies with a man I used to go toward Harvard Square in Cambridge is mad fun for
- 30: the first list was posted on the bulletin he brought home a turkey will die on my rug is deep with snow and sleet are destructive and playful students always
- 50: the next room to mine silver in Pennsylvania is late in getting home on time my date was tremendous fun going there skiing this day would end and have no more objections to his speech on the radio last night played the viola in the orchestra and chorus performed the

5-order approximation

- 10: they saw the play Saturday and sat down beside him
- 20: road in the country was insane especially in dreary rooms where they have some books to buy for studying Greek
- 30: go it will be pleasant to you when I am near the table in the dining room was crowded with people it crashed into were screaming that they had been
- 50: house to ask for is to earn our living by working towards a goal for his team in old New-York was a wonderful place wasn't it even pleasant to talk about and laugh hard when he tells lies he should not tell me the reason why you are is evident

7-order approximation

- 10: recognize her abilities in music after he scolded him before
- 20: easy if you know how to crochet you can make a simple scarf if they knew the color that it
- 30: won't do for the members what they most wanted in the course an interesting professor gave I went to at one o'clock stopped at his front door and rang the
- 50: then go ahead and do it if possible while I make an appointment I want to skip very much around the tree and back home again to eat dinner after the movie early so that we could get lunch because we liked her method for sewing blouses and skirts is

Text

- 10: the history of California is largely that of a railroad
- 20: more attention has been paid to diet but mostly in relation to disease and to the growth of young children
- 30: Archimedes was a lonely sort of eagle as a young man he had studied for a short time at Alexandria Egypt where he made a life-long friend a gifted mathematician
- 50: the old professor's seventieth birthday was made a great occasion for public honors and a gathering of his disciples and former pupils from all over Europe thereafter he lectured publicly less and less often and for ten years received a few of his students at his house near the university

[CHAPTER 6]

The Mass Media: Films, Tapes, and Television

Introduction

When Johnny's teacher flicks the switch that starts the movie projector, he is participating in a technological and social revolution in education of much wider dimensions than he realizes. When the lights go out, the teacher is no longer the focal point of attention; a machine and a film take over. After the lights go back on, the teacher may appear lackluster when compared to a film filled with dramatic visual content and accompanied by the stirring orchestration of an operatic overture.

If we add to the sound-track motion picture the full range of audio-visual stimulus materials and media available—television, teaching machines, language laboratories, recordings, slides, diagrams, models—we see that modern technology is already well established in the classroom. The problem confronting the teacher is how to master this array of materials and equipment, along with reading and writing, so that it may serve the educational purposes of the classroom and the school. The problem will not be easy to solve, because audio-visual "aids" are frequently not aids at all, but mass-produced robots in which few adjustments to local use can be made. For example, it seems at times that films decide to use the teacher and determine the content of the course. Thus teachers can easily allow audio-visual materials and media that are locally available to determine their teaching activities without considering what educational objectives are served. This is, truly, teaching in the dark.

The following articles raise several important questions about the use of audio-visual materials and media. Because there is such a rapid rate of increase in both knowledge and people, we have less time than ever before in which to educate. The concreteness and immediacy of audio-visual materials, in addition to their possible wide distribution, is an obvious way to begin tackling this problem. Many of the students in this country and abroad have not been born into a social or cultural tradition of learning that involves the printed page and fairly high levels of abstraction, a situation which is considered in the chapter on individual differences (pp. 479-526). Nor should we forget that in some areas of the world the first formal educational experiences of children will be with television and motion pictures and not with the printed page. We do not know now how this rather drastic alteration in the stimulus situation will modify their educational behavior.

Kendler and Sherburne both indicate that much investigation of a theoretical-practical nature should precede and accompany the use of audio-visual devices. In a new educational area, then, the student can see how the laboratory and classroom can be linked (see Chapter 1). One characteristic of the learning situation is motivation; some educators and psychologists have suggested that the chief use of films and television may be to arouse interest. One may grant that this does seem superior to Bugelski's Chinese water torture (see pp. 83-84), and that it avoids the aversive treatment against which Skinner protests. If the excitement of the media can lure children into responding to stimuli in their environment, we may have made an initial gain. However, this could degenerate into large-scale, low-budget school entertainment, with no significant gains in learning. At this point the student should review the evidence provided by Kersh, which indicates that greater motivation does not necessarily lead to greater learning (pp. 277-287).

We need research not only on the use of films, tapes, televised programs, etc., but also on their proper construction and on their effectiveness in accomplishing particular educational tasks *before* they are put on the market. It may be possible that films, sound tapes, and video-tapes could be made in the manner that standardized tests are made. We could then know for what purposes to use them, what kinds of results to expect, and with what particular

populations of students. This specification of student groups could help answer the question of individual differences, which becomes particularly difficult to answer when the schools turn to mass media (see Chapter 5). One form which the general research on audio-visual materials might take is excellently described by Kendler (pp. 384-393).

In reading the following selections, the student should keep in mind the various characteristics of learning situations which have been discussed in previous chapters and which are summarized by Gagné and Bolles (pp. 31-51). Does the use of films and television, for example, allow students to make the necessary responses? Does it guarantee adequate reinforcement? Does it permit the learning activity necessary in the "discovery" of principles?

Relationship of Readings in Chapter Six

Three readings in this section are general discussions of the use of mass media in the school and one a report of an experiment using educational television and large group instruction. Kendler suggests that the stimulus-response model used in experimental psychology might be applied to research on educational films. His paper is an excellent illustration of the translation of laboratory science into classroom use. The article by Sherburne is concerned with research on television and should be compared with that of Kendler. Sherburne is concerned with how the media itself can influence educational outcomes. He favors "internal research"; his view is in some ways consistent with Kendler's suggestion that we actually find out how learning occurs in the use of films. The article by Siegel and Macomber is an example of the research now carried on in educational television and in large-group instruction. The student can inspect the study to see if it meets the criteria for audio-visual research set by Kendler and Sherburne.

Stimulus-Response Psychology and Audiovisual Education *

The editor considers this to be a key article because it is a clear statement of the basic relationship which exists between experimental and educational psychology. It very much reflects the views of writers whose articles have appeared earlier in the book: Melton, Gagné and Bolles, McDonald, and Skinner. The author makes a convincing defense of the use of the stimulus-response (S-R) model in classroom research, and he removes some misconceptions which psychologists and laymen have about it. When the teacher, Mr. Jones, asks a question and Johnny raises his hand to answer it, we have a simple illustration of a stimulus-response relationship. In this example we can presume that the stimulus was Mr. Jones' question and the response was simply Johnny's correct answer to it. We can hope that Mr. Jones smiled or spoke his approval for Johnny's successful effort, because this would furnish the reinforcement necessary to keep that question (stimulus) linked to that answer (response).

Such a description of the stimulus-response relationship can be greatly embellished by considering and describing other internal, inferred, and mediating events which occur between the original stimulus and the final response and which keep them united. Kendler has given us a stripped-down model of learning while acknowledging that human learning is more complex. It would be interesting for the student to compare his views with those of preceding authors. For example, how does his view of the importance of the role of motivation and reinforcement compare with that of Skinner? Is it possible to explain the re-

* Reprinted with the permission of the author and publisher, the National Educational Association, from the article of the same title, *AV Communication Review*, 9, No. 5 (1961), 33-41. References are omitted.

search on cognitive learning in Chapter 4 in terms of the S-R model? If not, what additional research would be necessary before we could do so?

Using stimulus-response language Kendler discusses some practical problems which must be worked out in the use of audio-visual materials: (1) What advantages do audio-visual techniques have in presenting stimuli? How does he connect this with transfer of training? (2) What problems arise in obtaining and reinforcing correct responses? (3) Of the practical suggestions for eliciting responses, which would be the most effective? Why?

Anybody who is interested in controlling behavior—and every audiovisual educator is—must first learn to describe it. But, as the history of psychology demonstrates, this is no easy task. One look at raw behavior with its multitudinous facets will immediately raise doubts about any simple language system that aspires to describe behavior.

Psychologists have toiled with this problem since the birth of their science less than a century ago. Many descriptive systems have been created but few have survived. None have attained universal acceptance.

Nevertheless, there is one language system that has stood the test of time fairly well. And that system uses the stimulus-response conception of behavior. Unfortunately for students of behavior, there is still much confusion surrounding this S-R approach. The fault lies as much with its adherents as with its critics.

The proponents of S-R psychology must realize that the data and theories of learning do not *demand* an S-R analysis. The fact that a majority of psychologists who are concerned with learning have adopted an S-R approach testifies to its usefulness, but not to its intrinsic validity. The critics of S-R psychology must recognize that it consists of three separate components. First, it is a language that seeks to represent psychological events. Second, it is methodological orientation that is anchored in the behavioristic tradition of investigating psychological problems in an objective, experimental manner. Third, it is a classification used by many textbook writers

to group together different theories of learning (e.g., those of Guthrie, Hull, Spence, and Estes) that share a common language and methodological orientation.

Our major concern here is with the first and third components of S-R psychology. Behavioristic methodology has numerous philosophical ramifications that go well beyond the boundaries of a paper such as this. Nevertheless, it is helpful to mention that one of the distinctive features of the behavioristic approach is its passionate demand for *experimental* data to justify any interpretation of behavior, whether concerned with the responses of rats in mazes, or with the reactions of humans to a training film. Natural observation, clinical judgment, and personal experience all fall far short of the rules of evidence required by behaviorists to support any psychological explanation, whether couched in a modest guess or within a hypothesis embedded in an elegant theoretical system.

Many reservations have been expressed about describing behavior in terms of stimulus-response associations. Some of these stem from the apparent discrepancy between active, flowing behavior and the inert, static, single S-R association. It seems incomprehensible that raw behavior can be reduced to isolated S-R connections. This objection stems from a misunderstanding. To use S-R language does not mean that complex behavior *actually* consists of S-R connections. Relevant to this point is a quotation from Toulmin, an English philosopher of science, who after analyzing the concept of light, concludes, "We do not *find* light atomized into individual rays: we represent it as consisting of such rays." Applying the same idea to the concept of the S-R association results in the following statement, "We do not *find* behavior atomized into individual S-R associations: we represent it as consisting of such S-R associations." The concept of the S-R association, therefore, must be judged not in terms of its ability to provide a clear image of behavior, but rather in its capacity to represent the facts of behavior.

S-R language has proved its worth by the assistance it provides the experimental psychologist who seeks to translate his ideas or hunches into experimental problems from which reliable empirical knowledge can be obtained. This is evidenced by the fact that, during the last four decades, the S-R learning psychologists have been the most active experimental and theoretical group in psychology.

Their productivity is in some measure due to the fruitful and cleansing effect that stimulus-response language has upon designing, reporting, and interpreting research. This language forces the psychologist to focus his attention on objectively defined environmental and behavior variables, and thus encourages the collection of data and the testing of ideas.

The essence of the stimulus-response language is contained in the S-R paradigm, which functions as a *model of behavior*. Accordingly, there are three important sets of variables in psychology; stimuli, responses, and the association (the -) between them. The stimulus refers to some aspect of the environment, while the response points to some feature of the behavior. The hyphen represents the degree of relationship between the two. Whether or not the association is formed depends on whether it is reinforced. (The meaning of this term "reinforced" varies in different theoretical systems.)

Not all S-R psychologists are in complete agreement as to whether concepts in addition to those mentioned are needed to explain behavior. Many of them do add a motivational concept to their theoretical schema involving stimulus-response associations. Others believe that this addition is unnecessary. For them, the motivational concept is not basic since its effects can be reduced to the action of stimulus variables. In this paper, there is no need to get enmeshed in such theoretical disputes. When they are viewed from the distance of audiovisual education, most of these differences become attenuated, and some even disappear.

The reasons for the strong and persistent fascination S-R psychologists exhibit for the facts of conditioning should now be clear to the reader. Conditioning provides the clearest picture of how a response becomes associated with a stimulus, and how this connection can be strengthened or weakened. In this complicated world of behavior, conditioning provides the *simplest* form of learning. By observing it closely one can uncover the secrets of the learning process.

PRACTICAL APPLICATION

No doubt many of the more impatient readers are wondering how the above ivory-tower analysis is related to the practical problems of audiovisual education. The author perceives two important links.

forcing) conditions for learning to occur. The above description is of necessity brief and oversimplified. But the fundamental point is simply that the model of behavior involving stimulus, response, and associative processes has been a powerful tool in ordering the experimental data of learning and related phenomena. The belief expressed here is that if it is used with care and understanding, it will serve as an effective guide for the planning of audiovisual training programs. Fortunately for the educator, there are excellent secondary sources which describe in much more detail the meaning of the S-R model in a variety of situations and which analyze many of the subtleties that could not be touched upon in this paper.

DEVELOPMENTS IN LEARNING THEORY

About 30 years ago, when serious attempts to formulate theories of learning were begun, it seemed that the experimental work of the learning psychologists would expand continuously to include more and more complicated situations. Actually, a trend towards simplicity has operated. For example, complicated mazes once widely used to study the learning process of both animals and men have all but been abandoned. Psychologists found that variables associated with the particular maze pattern exerted such a powerful effect that the general principles governing behavior were hidden from view. The trend nowadays is to select experimental situations that highlight the influence of basic variables. The classically simple conditioning and discrimination learning situations (including the simplest maze pattern of all, the T) are more widely used than they were three decades ago. Complex problems are no longer used just to see how organisms respond to them. When the effect of new variables is to be evaluated, modifications are introduced into the basic learning tasks. Because these tasks have a known behavioral reference point, the effects of new variables can be judged precisely.

The strategy of limiting the number of experimental procedures in the psychology of learning to a few basic ones, has, I believe, a moral for research in audiovisual education. And this moral is that basic research procedures for audiovisual education must be developed. What is needed—for example, in studying learning which results from training films—is a group of brief sequences that can

function as standard experimental problems. Learning occurring in these situations could serve as a base from which the effects of variables in audiovisual education could be evaluated.

One reason why such research devices have not been developed is that the problem of evaluating training films has often been confused with research designed to understand the learning process in audiovisual education. Typically, most training films contain such a conglomeration of variables that it is impossible to isolate their individual effects. It should be understood that no criticism is being leveled against evaluation studies. They must be done. The point made here is that the evaluation of training films presents a different problem from that of understanding the audiovisual learning process. As such, they demand two different kinds of solutions.

In addition to the pressures towards simplifying experimental tasks, the recent history of learning theory exhibits another important trend. At one time, learning theorists thought it would be possible to integrate the facts of learning by relating independent environmental variables to the dependent behavior variables; they hoped to do so by postulating hypothetical intervening variables (theoretical concepts). It soon became apparent to them that this approach was not as simple as it seemed. They found that many of the independent variables selected did not exert a single influence. The effect of the variables depended on how the organism responded to them. For example, the influence of delaying the reward for the correct response was found to be not a function of the timing but, instead, of how the organism *behaved* during the delay.

Today, learning theorists are more cautious when they attempt to relate stimuli with overt responses. They want to know whether any other responses intervene between the two. They know that in a very simple conditioning situation it is possible to correlate the environmental stimulus with the response *it evokes*. But they have learned that to explain the behavior, particularly of humans, in more complex problems—such as discrimination learning, learning sets, and concept learning—it is often necessary to assume that mediating events are occurring between the external stimulus and the overt response.

An oversimplified version of a mediational mechanism assumes that the organism reacts to the external stimulus with a covert re-

sponse which produces an implicit cue that, in turn, elicits the overt response. The all-important feature of this mediational mechanism is that the behavior of the subject is controlled by a cue he himself emits. Therefore, the audiovisual educator interested in teaching a skill involving a mediational mechanism—for example, a verbal skill—must arrange a training sequence that elicits similar mediational responses among the various members of the audience. There are two major ways to do this. The first is to train the audience to make the necessary mediating responses so that later the correct overt response can be associated with its cues. The other alternative is to select an audience that can make the appropriate mediational responses, and to tailor the training sequence for them.

Once again we find ourselves back to the ticklish problem of encouraging the appropriate response by means of audiovisual training techniques. In an individual training situation where the cues are presented and the student is given an opportunity to make an overt response which can be rewarded when correct and left unrewarded when incorrect, the educational process is easy to control. When the responses are unknown and the reinforcements cannot be punctiliously administered, the problem becomes more complicated.

Some efforts within the traditional audiovisual educational program can be made to overcome these difficulties. One is to require the audience to make an overt response where it is feasible. Another possibility lies in instructing the audience as to what they should do during the film to facilitate training. Still another is to recognize the limitations of the audiovisual medium, and to supply the necessary training of the response in another situation. The final possibility may be the best of all. And that is to incorporate audiovisual techniques with those of auto-instructional devices—teaching machines or programed texts. The best of all education worlds will result from such an arrangement. The stimulus situation will be presented in the best possible manner. Simultaneously, the subject will have a chance to practice the correct response and to be reinforced for it.

SUMMARY

In summary, this paper has been concerned with the implications of stimulus-response psychology for audiovisual education. Initially, it was pointed out that stimulus-response psychology consists of

three separate components: a language system, a methodological orientation, and a group of theories. The language is useful in reducing the complexity of behavior to a set of manageable variables. By representing behavior in an S-R manner, it is possible to cope experimentally with, and ultimately to control it. The problem for the audiovisual educator becomes one of arranging the optimum conditions for the formation of associations between stimuli and responses. Some problems associated with this task were discussed. Particular emphasis was placed upon the strategy of devising fruitful experimental techniques and arranging for the evocation of the appropriate response. It was concluded that audiovisual educational techniques may achieve their greatest productivity when combined with the techniques of teaching machines.

References

- DEESE, J., *The Psychology of Learning*. Second edition. New York: McGraw-Hill, 1958.
- HULL, C. L., *The Principles of Behavior*. New York: Appleton-Century-Crofts, 1943.
- KIMBLE, G. A., *Hilgard and Marquis' Conditioning and Learning*. Second edition. New York: Appleton-Century-Crofts, 1961.
- ROSHAL, S. M., *The Effectiveness of Film Mediated Perceptual-Motor Learning with Varying Presentation of the Task*. Ph.D. Thesis, Pennsylvania State College, 1949. University Microfilm, Ann Arbor, Michigan.
- SKINNER, B. F., "Teaching Machines." *Science* 128:969-77; October 1958.
- TOULMIN, S. E., *The Philosophy of Science*. London: Hutchinson University Library, 1953.

into the doing, the measuring, and the communicating about the doing and the measuring. There is no place left for thinking about what we have done or where we are going.

Generalized, this sounds obvious. Let us examine in more detail how this applies to the educational television scene. In the following pages, I will attempt to show that in our speedy and unthinking embrace of this new technique, it has been stereotyped as a solution to money and number problems in education, and that there has been an almost complete disregard for first questions, basic ones about the nature of this truly new medium. For certain biases have bent the twig unduly, and influenced the course of experimentation and research. And here, as in any situation, the questions formulated predetermine the answers we obtain, and strongly influence the image of educational television which results.

Before pinpointing what is wrong with the image of educational television or suggesting how it may be broadened, there is reason to review what the present picture is. Using as our basis the research and predominant activities within the field, how is "educational television" perceived?

Primarily, educational television has its widest support as an electronic extension of classroom-style teaching of the "best teacher" to the "most" students, generally for formalized education, and justified most often on a purely economic basis.

True, in the early days there was much talk of discovering dynamic ways of using television, and of bringing mass cultural efforts to the masses out of school. But today, with over 40 stations on the air and hundreds of schools using television, close examination of the kind of broadcasting, the sources of support, and the growth in closed-circuit systems reveals a concept of TV as a device primarily of educational expediency.

A detailed study or a glance at the reports and studies done in the past ten years reveals more. There one can see the kind of questions asked. For rather than trying to find out how to improve the quality and enlarge our vision of imparting information with this new and valuable tool, we have proved over and over again that television can extend the teacher beyond the limit of one moment and one place, and still teach. An obvious conclusion but, for emphasis, let us examine its ramifications.

The first major study of teaching by television occurred in 1949, and engaged three Naval Air Stations in an evaluation of TV for mass training. The Navy asked, "Do men taught by television learn as much as men taught by conventional instruction?" And the reply, a familiar one: ". . . 80 percent of comparisons showed television as good or better than conventional classroom instruction." In the ten years since this study, the question has been amplified, modified, restated, and extended in its application to a multitude of educational situations.

A perusal of the AVCR Research Abstracts, the NAEB Fact Sheets, or any other research summaries, shows the enormous repetition of this question. And as a result we now know that television can be used to teach high school students, elementary pupils, the Army, the Navy, college students, housewives, student teachers, and IBM salesmen. It can act as a conveyor of learning for subjects including home nursing, physics, algebra, dress making, college composition, typewriting, psychology, and maintenance of military telephones. It can do this, amazingly enough, in Florida, Alabama, Pennsylvania, Illinois, California, and even Canada, England, and France. And apparently students can learn at 6:30 A.M., noon, late afternoon, the early evening, or late at night. We have no examples as to the effects of high altitudes, teaching in Urdu, or the applications to pinball machine maintenance, but these will undoubtedly be forthcoming since research projects are multiplying like rabbits throughout the country (and I might add, just as indiscriminately).

True, a number of other questions have been asked and researched, but they are of minor importance since they are infrequent in occurrence, and the results are little regarded. In most cases, even these different questions are still offshoots of the old "can television mass-teach" query, and thus have little over-all effect on enlarging a concept of the nature of television in education.

In trying to clarify and extend our thinking I do not wish to imply that all of the past research has been useless, nor do I wish to diminish the importance of television as an electronic tool to extend the formalized educational situation. This, educational television is, most certainly. It has applications of tremendous import to problems of mass education, overcrowded classrooms, overloaded teaching schedules and overextended budgets. But television used in

education can be much more. It offers an opportunity to teach better, to reorganize material for faster and for different kinds of comprehension. It may actually affect the ordering of knowledge itself, and not merely multiply and transmit more widely our old manners and mistakes.

To amplify, let me introduce two terms to describe the kinds of investigations we are and are not doing. They are: *internal research* and *external research*.

Internal research is concerned with factors occurring within the television presentation. Primary concern here is not with what happens to the viewer, but how the material is brought together, organized, presented, and made a complete visual-aural unit. The questions inherent here are fundamental ones of learning and teaching—organization for visual, aural presentation, idea density, word level, the television teacher, visual practices, relationship of the visual to the aural, and many other questions basic to learning and teaching through a medium that enables one to "streamline" the message.

External research involves questions external to the presentation. Here one is concerned with effect, retention of subject matter, comparisons with traditional teaching, faculty and student attitudes, administrative problems, costs, techniques and mechanics of integrating television into the curriculum, the physical set-up for viewing, and many other factors.

Both are interrelated, but there is an important distinction between the two kinds of concerns. To extend the illustration let us apply it to printing, a medium so much a part of our culture that we have an unconscious understanding of its nature.

Concern with external factors in printing would find research on questions such as whether people can learn from reading, who reads books, the nature and cost of publishing, testing of reading to learn whether it communicates in lower grades as well as on upper levels, and comparing reading a lesson to hearing it spoken to see if learning occurs at both times.

Internal concerns would be questions on the nature of communication by writing, how ideas are best conveyed by words, what constitutes good writing, how illustrations enhance the writing, how writing for books is accomplished, and so on.

While obvious, such a comparison helps to delineate the situation. We in educational television have been mostly measuring our tool, not giving a whit about understanding its nature.

If one examines the research of the past ten years one sees that a large percentage falls into the "external" area. This trend seems to be continuing in the grants under Title VII of the National Defense Education Act. Out of 31 TV proposals approved within the past year, one can make a rough analysis of the kind of research studies that are currently being supported. Some are difficult to classify (by the brief titles provided) but it seems clear that 26 of the 31 are concerned with external factors. In other words, 84 percent of the research is aimed at measuring factors external to the presentation itself. Only 5 (16 percent) examine internal factors.

What specifically are the kinds of questions and research which will be engendered? I obviously cannot predict with any accuracy, but I can say what kind of questions and research I hope will be undertaken in the decade to come.

The next major effort in educational television research should arise out of the complex of questions which considers what the nature of television's pictorial-verbal communication really is. In short, the emphasis in ETV research should change from "external" to "internal."

This is important because in much of our research, we have been acting as though the content of TV and the content of classroom instruction are essentially the same. And yet we do know that there are some major differences between what is communicated by television and what is communicated in other ways—or at least, that there can be differences.

In a study by Edmund Carpenter of the University of Toronto, and reported in *Explorations VII*,¹ he pointed out that the medium of communication affects the content to such an extent that it can really be considered a part of the content. Or, "the envelope is a part of the message." In other words, if a medium is to be used to its best advantage, it will "say" different things from what would be "said" in another medium. As Carpenter comments, the content is

¹ Edmund Carpenter, "The New Languages," *Explorations VII*, University of Toronto, Toronto, Canada, March 1957.

not a sheet of paper which can be transferred from the envelope of one medium to another without modification. A change in medium is a change in content.

Such investigation into the pictorial-verbal nature of communication has not been thoroughly explored, even in "instructional" films. And in TV, we are still trying to find pat formulas for making material appealing and different by omitting blackboards, and leaving out the impedimenta of old style "teaching"—this, instead of finding concepts of how to use television through understanding of the medium and its effect on content.

Under this pictorial-verbal study there might be at least two general areas of concern: *potentialities* and *pedagogy*.

The first, potentialities, involves us in studies which will provide academic description and clarification of what television can do best, and cannot do. It will provide information about the nature of TV communication, and will give us some insight into what we are communicating.

Let me be specific. For the sake of familiarity let us take a poet and a poem, both well-known: Robert Frost and "Birches." Without extensive thinking, one can visualize six very different types of experiences which could be presented by television relating to this material. For example: (1) An excellent teacher reads and discusses the poem and leads in an analysis of its content. (2) New England birch country and seasons illustrate the poem through film. (3) Robert Frost reads and converses about his poem. (4) A panel of students discuss the poem with a teacher. (5) An actor reads the poem with dramatic emphasis. (6) The actual words of the poem are shown on the screen and analyzed as to word choice, rhythm, length of line.

None of these are unusual. It is only what we do not know about the nature of the various experiences involved which is interesting. What happens to a student who sees and hears Frost read his own poem as opposed to the student who sees the actor reading the poem (not "which teaches better," but "what is the nature of the differences in what is learned?"). What is the nature of the concepts communicated in the various approaches? What can we communicate by television that we cannot communicate by other means? What should be learned by picture alone, by sound alone, by picture and sound? What stimulates intellectual curiosity, empathy, motivation

to understand? And how does television as the "envelope" enhance, impede, or activate the message to produce these? All such questions go beyond whether the student learns as well as by the conventional classroom method. They attempt to uncover information about what has been thought of as elusive configurations of learning. Television, certainly, did not raise all these questions but it may enable us to answer them, for recorded teaching experiences at least make it possible for us to compare and test.

When some basic understanding is at hand as to television's distinctions from and likenesses to film, books, comic strips, then another of our tasks will be easier. I refer to the establishment of some reasonable criteria of success in presentations by television. At the present time the "successful" ETV programs are those which receive "awards" or boast larger audiences than the others. Is it not unfortunate that those who are "educating" accept as their criteria for achievement the standards which assist an advertiser in buying commercial time?

If you wish to check on the state of analysis in educational television try criticizing a program to any one of your colleagues. Most likely you will hear, "well so and so likes it," or "it was the most popular of . . .," with no attempt to meet the issue directly. Our failure to criticize is born out of a lack of criteria which we understand and can bring to bear on the issues. Perhaps an exploration into the nature of TV can help us.

The second area of research would come under the general rubric of pedagogy. I refer mainly to the television teacher. When selection is made of a teacher for a local video classroom or for Continental Classroom the best we can do is to try to find someone who is by common agreement an excellent teacher.

But what does this mean? What characterizes these so-called "good" teachers? Is a good classroom teacher automatically a success on television? What are the qualities enabling a teacher to use television effectively? What are the characteristics of a person who can meet the demanding role of a performer and teacher? While television did not create this question of "what is a good teacher?" (as it did not many of the other major ones), it may enable us to help answer it. For the first time in the history of communication we can have a record of teaching—hours of it by variously skilled teachers. These can be studied, subjectively and objectively, for definable

qualities. When refined teaching skills are amplified by the uniqueness of television then we will really have scored in communication.

When we have answered some of the questions about the nature of television's pictorial-verbal communication, we may well return to more mature consideration of the external aspects. I mean when we better know television and what we can "say" with it, then we will have a better idea of what we ought to be accomplishing. External research can be rejuvenated and the answers from this aspect of research can be made more significant.

I have said that television research needs to escape the redundancy of proving it can teach and that it should move on to more basic questions. I have explained why I think that we have been pedestrian in our present research, and I have suggested the kind of biases which I feel have affected thinking in this area. Lastly, I have suggested the kinds of research which I feel we should undertake in the next decade.

Let me say that I do not feel that by doing this, we will have solved all of our problems. In fact, I would warn you that if these things are accomplished, then will be the time to be brave. For we may discover things which will revolutionize some of our designs for learning. Some subjects may best be visualized. Others will not warrant the expense of a TV tube, for they will be the kind which can only be taught through exchange between teacher and pupil. Television explored may force us to revise our concepts of how rapidly subjects can be taught. Some subjects, we may find, can be taught more rapidly and efficiently than now, where others will require as much or even more time because of the great amount of subject matter which can be made readily available to the student. The knowledge itself which we are passing on to succeeding generations may even be of a different nature, since its ordering may have a new, pictorial-verbal emphasis rather than the purely verbal one of today.

And television's most important contribution may turn out to be not the solving of the crisis in numbers to be taught, but an attack on the crisis which is not yet so apparent. I mean the crisis of rapidly expanding knowledge. The day is past when a nation can survive and leave its books (or its educational TV programs) behind after graduation. And the day may arrive when everyone will enroll in a continuing university carried on by television, with an un-

precedented teaching efficiency made possible by studies of pictorial-verbal communication. My thesis has been general on purpose. For unless in the decade ahead we broaden our image of television to be something more than just an electronic extension of what already is, then truly we can be ashamed. For it would mean that we have not been brave enough or creative enough to ask the difficult questions, and having answered them, to take advantage of what those answers offer.

References

- CARPENTER, EDMUND, "The New Languages." *Explorations VII*. University of Toronto, Toronto, Canada, March 1957.
- Title VII—News and Reports*. Educational Media Branch, U.S. Office of Education, Washington, D.C., June 1959, April 1959, and February 1960. [See also *News and Reports* for July 1960 and recent listing page 222 this issue.—The Ed.]
- Training by Television*, SDC Report 476-02-2. Special Devices Center, Port Washington, New York (undated).

LAURENCE SIEGEL and F. G. MACOMBER
Miami University, Ohio

Comparative Effectiveness of Televised and Large Classes and of Small Sections*

The following report, an outstanding example of research, compares the effectiveness of televised and large-group instruction with more conventional instruction in small classes. For rea-

* Reprinted with the permission of the author and the American Psychological Association from the article of the same title, *Journal of Educational Psychology*, 48 (1957), 371-382. Footnote and references are omitted. The *Final Report* is now published.

sions that are not altogether clear, a class of about thirty students and one teacher is supposed to be the ideal size for a good teaching situation. To raise the size to forty or more, as increasing enrollments frequently force us to do, is, supposedly, to sacrifice all possibility of successfully teaching the class. When one presses the defenders of the small class for a psychological defense of their position, there is frequently a retreat to clinical language; one hears allusions to "rapport," "empathy," "the basic uniqueness of each living being," and the "need to *know* one's students." It may be granted that these are all terms which have merit, but they involve so much more that they are hardly useful in an objective discussion of the issue. The reports of research on this question rarely ascribe to groups of either size an advantage in learning efficiency.

The following report permits the student to test his research know-how. To guide his efforts, these questions should be considered: (1) In what way does this research fit the characteristics Sherburne describes for "external" and "internal" research (pp. 398-400)? (2) The experiment is not planned with Kendler's stimulus-response (S-R) model in mind. What changes in Siegel's experiment could be made to make it conform more closely to Kendler's model? (3) What definition of achievement in the course were the investigators using? Why did they feel this was not adequate? (4) What various explanations could account for the reports of dissatisfaction from the students in the large groups and televised section? (5) Why might the high-ability students feel dissatisfied? (6) Under what circumstances would it be possible for students to express dissatisfaction with instruction in large groups which are actually promoting efficient learning?

College enrollments for the next decade are expected to rise sharply with an anticipated shortage of well-qualified college teachers in most subject areas. These conditions will place additional pressures upon public institutions to increase class size. The development of educational television, and particularly of relatively low-cost

closed-circuit systems, raises the interesting possibility of teaching large groups of students while overcoming some of the visual difficulties inherent in large class instruction.

Miami University is engaged in a research program directed toward the comparison of the relative effectiveness of large group instructional procedures with smaller conventional-type classes of approximately 30 to 35 students. Two of the general approaches to large group instruction under investigation are designated LC (large classes with direct visual contact between student and instructor) and TV (closed-circuit television) classes. This paper summarizes the major results obtained during the Spring Semester, 1956. Three problem areas were investigated:

1. Achievement as a function of assignment to experimental or control sections;
2. Students' attitudes about the instructor and course as a function of section assignment;
3. Students' end-of-the-semester attitudes about having received TV or LC instruction rather than conventional instruction.

PROCEDURE

Subjects. The first two problem areas under investigation required the comparison of results obtained from experimental and control sections of the eight undergraduate courses listed in Table 1. Each experimental section was taught either as a large class (primarily lecture) or by means of closed-circuit television. The control sections were taught by conventional procedures. Both experimental and control sections were taught by the same instructor and no known biasing factor was known to be operative as a determinant of student registration for the experimental or control section. Nevertheless, the relatively small sizes of the control groups prohibited sole dependence upon random assignment of students to balance the effects of extraneous variables. Consequently, the experimental and control groups for each course were deliberately equated on the basis of scores on the Cooperative English Test (Form T), Cooperative Mathematics Achievement Test (Form P), and total score on the American Council on Education Examination (1948 edition), as well as grade-point-average for the previous semester.

The matching procedure required that some of the students actually enrolled in the experimental and control sections of each course be excluded from the specific experimental and control groups under investigation. All decisions about the constituency of the experimental and control groups were made prior to the collection of research data. Neither the instructors nor the students were informed regarding the actual composition of the groups providing data for the Experimental Study in Instructional Procedures. The total enrollments in each course and the numbers of subjects included in the study are summarized in Table 1.

Table 1
TOTAL ENROLLMENT AND NUMBERS OF EXPERIMENTAL
SUBJECTS IN EACH COURSE

Course Title	Number of Sections			Number of Students Enrolled in Sections ^a			Number of Subjects from Each Section ^b		
	TV	LC	Control	TV	LC	Control	TV	LC	Control
Foundations of Human Behavior	1		2 ^c	103		74	97		62
Principles of Human Physiology	1		1	96		31	90		30
Introductory Sociology	1		4 ^d	141		144	131		129
Human Biology	1	1		183	178		183	177	
Business and Government		1	1		119	35		105	30
Composition and Literature		2	2 ^c		100	50		37	37
Essentials of Modn. Geography		3	3 ^e		213	93		88	85
Elementary Psychology		1	1		78	30		77	30

^a Enrollments in multiple sections were divided approximately equally between each of the sections.

^b These are the numbers of students from each section actually providing data for the Study.

^c Multiple sections taught by the same instructor.

^d Multiple sections taught by different instructors, each of whom rotated TV presentations.

^e Multiple sections taught by different instructors, each of whom also taught an experimental section.

A full description of the matching data is given elsewhere. It is sufficient here to indicate that none of the mean differences between the experimental and control groups in any course on any of the matching variables was statistically significant. Equating of the experimental and control groups on the critical variables of ACE and grade-point-average was further supported by the lack of statistical significance of the chi squares between pairs of score distributions on these two variables. Furthermore, both groups in each of the eight courses were demonstrated to be equated both with respect to proportional sex distributions and proportional distributions by class standing.

Teaching Procedures. The specific instructional procedures within each course were not standardized. Rather, instructors were encouraged to develop those procedures which seemed best suited to the instructional situation, and which were best calculated to achieve the objectives of the course. Thus there were considerable differences in the manner in which the four televised courses were conducted. The Foundations of Human Behavior class met for two weekly periods of 90 minutes each; Sociology and Human Biology were convened for three weekly periods of 50 minutes each. The students in the experimental Physiology section met for three 50-minute televised presentations and for a two-hour laboratory period each week. The nature of the televised presentations also differed in the four courses. In both Biology and Physiology, the general method was that of lecture supplemented by numerous demonstrations and training aids. Five different members of the department participated in the televised presentations of Sociology. These presentations were approximately 30 minutes in length with the balance of the 50-minute period devoted to a discussion in each receiving room under the leadership of a regular staff member. In Foundations of Human Behavior, the presentations took a variety of forms, but with approximately one-third of each class period given over to discussion under the leadership of graduate assistants who served as monitors in each of the receiving rooms.

Similarly, there was no standard form of procedure in the large classes not utilizing television as a medium of presentation. The LC sections of English followed essentially the same procedure as that designed for the control sections except that only half as many

written themes were required, and, of course, individual student participation was considerably limited by the size of the experimental groups. In Geography, the students received two lectures a week with a third period devoted to small group discussion. The remaining two LC courses were given as three 50-minute lectures per week.

The control sections of all courses were conducted as discussion groups with lectures and demonstrations inserted by the instructor whenever he thought it desirable.

This diversity of approach was believed desirable even though it prohibited direct comparisons between courses. Forcing the instructors to adhere to a specific and inflexible type of presentation in both their experimental and control sections would have prevented them from capitalizing upon the advantages inherent in the specific classroom situation and would have emphasized some of the deficiencies.

Evaluative Instruments. Two classes of instruments provided data for this series of investigations: achievement tests and attitudinal measures. These two classes of instruments are described separately below.

"Achievement" was operationally defined for the present sequence of studies as performance on the objective portions of course examinations administered for the purpose of assigning final grades. It is recognized that such examinations typically measure only a single dimension of achievement, i.e., subject-matter knowledge. Thus the present data are not germane to such important course objectives as "synthesis," "problem solution" and "critical thinking." The possibility of leakage of information about the test was overcome by simultaneous administration of each examination to the experimental and control sections of each course with one exception. (Simultaneous administration in Biology was impossible because the experimental and control sections were taught during successive semesters.) Estimates of reliability of each test (Kuder-Richardson Formula 21) ranged from .55 to .93 with a median of .88. To the extent that the assumptions implicit in this computation were not met, the resultant values are properly underestimates.

Four Thurstone-type scales were included in the attitude battery administered near the end of the semester to students in experimental and control sections. These scales yield quantifiable informa-

tion about (a) students' appraisal of the effectiveness of the instructor in presenting material and handling the class (the I scale); (b) students' appraisal of the course content (the C scale); (c) students' attitudes about having received televised rather than conventional instruction (the TV scale); and (d) students' attitudes about having received large class instruction rather than conventional instruction (the LC scale). Each of these scales provided a nine-point continuum with 5.0 as the neutral position. Student judges were requested to rate the items included in a preliminary pool, and items were selected for inclusion in the final forms of these instruments on the basis of low Q value and the median of the distribution of judgments.

Split-half reliabilities (corrected for length of test) based on responses of a sample of 300 respondents were .91 for the I scale and .92 for the C scale. The intercorrelation between these two attitude scales was .52. The corrected split-half reliability for the TV scale was .89 and for the LC scale was .92 in a sample of 100 students.

RESULTS

Conclusions based on the results obtained from a single semester's instruction are subject to the influence of a variety of sampling errors. Only four TV and four LC courses were investigated. Conclusions based on these data might not apply to other courses or even to the same content taught by other instructors. Also, results obtained from the present samples of students may not generalize to other samples enrolled in the same courses chiefly because of the relatively small size of some of the control groups.

Achievement. The distributions of achievement test scores in the experimental and control sections as well as appropriate *t* ratios are summarized in Table 2. The null hypothesis is not refuted for any course except Biology, i.e., with the exception of this one course, comparable achievement resulted from experimental and control presentations.

Interpretation of the statistically significant finding in the case of Biology is confounded by the fact that the experimental and control sections of this course were given during successive semesters rather than simultaneously. Although a definitive interpretation of the results obtained for Biology is impossible at the present

Table 2
COMPARATIVE ACHIEVEMENT IN EXPERIMENTAL AND
CONTROL SECTIONS

Course	Section	TV Phase Achievement			t ratio	Course	Section	LC Phase Achievement			t ratio
		M		SD				M		SD	
Foundations of Human Behavior	TV	123.76	16.03		0.25	Business and Government	LC	70.03	8.58		0.70
	Control	123.11	15.05				Control	71.28	7.65		
Introductory Sociology	TV	110.09	20.91		0.24	Composition and Literature	LC	28.94	5.50		1.33
	Control	109.53	16.90				Control	30.78	4.74		
Physiology	TV	208.92	31.33		0.66	Essentials of Geography	LC	79.83	10.50		0.83
	Control	204.59	26.44				Control	78.54	9.56		
Human Biology	TV	253.97	15.21		3.00*	Elementary Psychology	LC	100.45	15.93		1.77
	Control	248.66	15.76				Control	102.52	21.77		

* Significant at better than the .001 level.

time, it is apparent that neither TV nor LC presentations adversely affected achievement as defined by subject matter examinations. This finding for TV courses is in agreement with results reported earlier by the Pennsylvania State University.

The absence of differential achievement in experimental and control sections does not negate the possibility of differential achievement as a function of academic ability in interaction with section assignment. In order to investigate the possible existence of such interaction, the students within each experimental and control group were dichotomized with respect to ability on the basis of percentile conversion of total ACE score. The cutting point for this dichotomization was the fiftieth percentile. Comparative achievement of the "high ability" subgroups in the experimental and control sections and of the "low ability" subgroups in these sections is summarized in Table 3.

The data presented in Table 3 support the conclusion that level of ability does not interact with assignment to a TV section or its control counterpart as a dual determinant of achievement. A similar conclusion applies to all LC courses with the exception of Composition and Literature. The statistically significant difference for the

Table 3

COMPARATIVE ACHIEVEMENT AS A DUAL FUNCTION OF SECTION
ASSIGNMENT AND LEVEL OF ABILITY

Course	Section	High Ability Subgroup				Low Ability Subgroup			
		Achievement Test			<i>t</i> ratio	Achievement Test			<i>t</i> ratio
		<i>M</i>	<i>SD</i>	<i>N</i> ^a		<i>M</i>	<i>SD</i>	<i>N</i> ^a	
Foundations of Human Behavior	TV	126.28	29.74	43	0.34	115.81	16.06	43	1.37
	Control	128.38	13.92	26		120.96	14.77	29	
Introductory Sociology	TV	117.80	13.13	55	0.97	103.04	14.67	46	0.12
	Control	115.52	12.60	67		102.70	9.54	40	
Physiology	TV	221.70	29.34	37	0.41	197.87	29.20	39	0.25
	Control	217.73	19.27	11		195.76	27.28	17	
Human Biology	TV	258.03	13.81	95	2.20 ^c	249.56	18.84	86	2.61 ^c
	Control	253.14	16.50	98		241.45	17.12	68	
Business and Government	LC	70.86	8.06	49	0.81	69.02	6.93	51	0.31
	Control	72.86	8.14	14		69.80	8.89	15	
Composition and Literature	LC	32.43	2.79	14	0.06	26.73	5.57	22	2.09 ^b
	Control	32.33	5.29	15		29.73	3.95	22	
Essentials of Modn. Geography	LC	83.81	9.19	42	1.14	76.11	10.49	43	0.22
	Control	81.59	8.32	41		75.63	9.80	44	
Elementary Psychology	LC	113.02	13.98	44	1.03	104.53	17.15	32	1.75
	Control	108.61	17.69	18		92.55	24.06	11	

^a Fluctuations from total *N* in the sample caused by unavailability of ACE scores or absence from final examination.

^b $p < .05$.

^c $p < .02$.

"low ability" students in this one instance is the result of both a mean difference and a variance difference. It might have resulted from the unreliability of the final examination or from sampling errors. If replicated with a more reliable criterion and larger samples, it would indicate that this is one course wherein the less able students profit more from small class instruction than from LC presentations. The high ability students in this course, however, performed equally well under both classroom conditions.

Students' Attitudes about the Instructor and Course. It is desirable to camouflage the identity of specific courses and instructors

for presentation of attitudinal data. These data are presented for courses designated by a letter rather than a title. The order in which the courses are hereafter listed differs from the order of listing in the previous discussions.

Group data summarizing the results obtained from administration of the I scale and the C scale in the experimental and control sections are summarized in Table 5. In courses taught by multiple instructors, the mean scores on both scales were obtained by pooling the data across instructors.

It is apparent from the data exhibited in Table 4 that TV instructors were rated at least as favorably by students in their television sections as by students in their control sections. Two of the four LC instructors, however, received significantly less favorable ratings in the experimental section than in the control section.

Table 4
SUMMARY OF I-SCALE AND C-SCALE RATINGS IN EXPERIMENTAL
AND CONTROL SECTIONS OF EACH COURSE

Course	Section	I Scale ^a			C Scale ^b		
		<i>M</i>	<i>SD</i>	<i>t</i> ratio	<i>M</i>	<i>SD</i>	<i>t</i> ratio
A	TV	2.80	1.00	—	3.08	0.50	—
	Control ^b	—	—		—	—	
B	TV	4.56	1.74	1.35	4.50	1.24	2.27*
	Control	4.10	1.40		4.00	0.91	
C	TV	4.16	0.88	2.86**	4.39	1.26	0.94
	Control	4.56	1.24		4.23	1.26	
D	TV	2.94	1.02	0.11	3.43	0.88	0.94
	Control	3.14	1.00		3.28	0.85	
E	LC	4.64	1.16	2.14*	3.91	1.07	3.18**
	Control	4.00	1.48		3.21	0.86	
F	LC	3.58	1.06	2.67**	3.83	1.07	2.88**
	Control	3.10	1.16		3.37	0.89	
G	LC	3.42	1.08	0.77	3.97	1.10	1.26
	Control	3.22	0.98		3.68	0.83	
H	LC	3.60	0.80	0.00	3.93	0.95	1.90
	Control	3.62	0.40		3.50	0.95	

^a Scores below 5.0 denote ratings as "more effective than average"; the lower the mean score, the more favorable the rating.

^b Data not available.

* Significant between the .05 and .01 levels.

** Significant at better than the .01 level.

Referring again to Table 4, there is a pronounced tendency for students in the control sections to rate the course content more favorably than students in either the TV or LC sections, although the obtained difference is statistically significant in only three of the seven courses for which data were available. Apparently, instruction given simultaneously to large groups of students (either by TV or in large classes) is accomplished at some sacrifice of the students' favorability to the course even though it does not affect their examination performance. Although the data are not amenable to direct comparison of TV and LC instruction, they lead to the suspicion that this deficiency may be more marked in the case of LC instruction than it is for TV instruction.

Attitudes about TV and LC Instruction. Group responses to the TV and LC scales are summarized in Table 5. The fiducial limits

Table 5
SUMMARY OF TV AND LC SCALE RESPONSES

TV Scale				LC Scale			
Course	<i>M</i>	<i>SD</i>	Fiducial Limits*	Course	<i>M</i>	<i>SD</i>	Fiducial Limits*
A	4.91	0.83	±0.14	E	5.74	0.72	±0.16
B	6.19	0.81	±0.18	F	5.45	0.84	±0.19
C	5.42	0.99	±0.20	G	5.35	0.81	±0.26
D	4.76	0.94	±0.20	H	6.03	0.64	±0.13

* Fiducial limits of the mean calculated for the .05 level of confidence.

extending the mean allow for an absolute interpretation of this value in terms of favorability or unfavorability to TV and LC instruction (taking 5.0 as a neutral position).

Two conclusions follow from these data:

1. It is impossible to generalize on students' attitudes about TV instruction. Students in one course preferred TV instruction to conventional instruction (Course D); those in another course considered the two types of presentation to be of approximately equal effectiveness (Course A). However, students in two of the TV courses (B and C) felt that television was inferior to conventional instruction.

2. Students in all LC courses would have preferred assignment to

a small section. The degree of this unfavorability, however, varied from slight (Courses F and G) to considerable (Course H).

The apparent variability in TV and LC attitudes from course to course is indicative of the fact that factors other than section assignment play a role in conditioning students' attitudes about the mode of instruction. The potential influence of two such factors (students' ratings of the instructor and academic ability) was investigated. Correlations between both I scale responses and ACE total score with responses to the TV and LC scales are exhibited in Table 6. Note that significant correlations involving the ACE are

Table 6

CORRELATIONS BETWEEN I-SCALE RESPONSES AND ACE TOTAL SCORE WITH
ATTITUDE TOWARD EXPERIMENTAL TYPES OF INSTRUCTION

Course	TV Phase		Course	LC Phase	
	I Scale	ACE		I Scale	ACE
A	.25 ^b	.00	E	.32 ^b	.06
B	.10	.32 ^b	F	.13	.02
C	.17	.23 ^a	G	.10	.38 ^a
D	.36 ^b	.14	H	.24 ^a	.04

^a r exceeds 1.96 times standard error of r of 0.00 ($p < .05$).

^b r exceeds 2.58 times standard error of r of 0.00 ($p < .01$).

indicative of an inverse relationship, i.e., the higher the ACE score, the less favorable the TV or LC attitude.

It is apparent that if any ability subgroup is unfavorable to TV or LC instruction, it is more likely to be the high ability students than the low ability students. However, the inverse relationship between academic ability and attitudes toward the mode of presentation may be overcome by factors which are not yet entirely clear. One such mitigating factor is probably student perception about the instructor. There is a marked tendency for academic ability to correlate inversely with TV and LC attitudes in those courses where ratings of the instructor do *not* correlate with these attitudes.

CONCLUSIONS

1. Acquisition of subject-matter knowledge as measured by objectivity retest performance was not adversely affected by assignment

to large (experimental) section (TV or LC) rather than to a small (control) section.

2. Academic ability did not interact with type of section assignment (experimental or control) as a complex determinant of achievement in any of the televised courses. A similar conclusion applies also to the LC courses under investigation with the possible exception of Composition and Literature.

3. Student ratings of the instructor (I scale) were not adversely affected by television. The only statistically significant difference between ratings by the TV and control sections occurred in one course wherein the instructor was rated more favorably by his TV section. Ratings of the instructor did tend, however, to be less favorable in LC sections than in control sections. The mean difference was statistically significant in two of the experimental courses.

4. Student ratings of the course (C scale) were less favorable in both TV and LC sections than in the control sections. This deficiency appears to be even more serious in LC sections than in TV sections (although data amenable to a direct comparison between these two types of presentation were not available).

5. Students' attitudes about TV instruction were not uniformly held across courses. In contrast, students in all LC courses were unfavorable to the mode of instruction when compared to small class instruction.

6. When academic ability was related at all to students' TV and LC attitudes, the direction of this relationship was inverse, i.e., high ability students were less favorable than low ability students. However, the fact that such a relationship was not present in all courses suggests the likelihood that other factors interact with academic ability in this regard. One such interacting factor is probably the students' attitudes about the effectiveness of the instructor.

References

- CARPENTER, C. R., & GREENHILL, L. P., *An investigation of television for teaching university courses*. University Park: Pennsylvania State Univer., 1956.
- MACOMBER, F. G., & SIEGEL, L., *Progress report of the experimental study in instructional procedures*. Oxford: Miami Univer., 1956.

Is Education Enough?

Introduction

We have often been reluctant to fully face the fact that in the American public school many students are much more or much less intellectually able than many other students. Although progressive education began as a protest against inflexible school practices and against teachers who were insensitive to the unique personal qualities of each child and to his limitations, little attention was given to the fact that many children were particularly bright and others perplexingly dull. After all, progressive educationists were more interested in improving the emotional than the intellectual climate of the school. They also were part of a larger reform movement that tried to improve the social and economic lot of the underprivileged child. They did not want the children of the poor, whatever their ability, shut out of the school. The tenor of the times has strikingly changed in the last decade; now we can openly discuss the problems which arise in the school because of differences in intelligence.

Even this book of readings, which focuses on the problem of producing more efficient learning in the classroom, could contribute to Utopian educational dreams. In fact, the chapter from Skinner's novel, *Walden Two*, was included as a deliberate step in that direction. It remains the editor's passionate belief that much more can be discovered about the characteristics of practical learning situations, and about their management, than we now know. We could avoid much of the waste of present misguided effort. Until now, however, the discussion has been carried on in a biological vacuum; we have not attended to the underlying physical organism in which all this learning is supposedly taking place. If we separate learning too far

from the organism in which it is occurring, we may be guilty of a peculiarly modern version of the mind-body dualism. We could even go along with Skinner and ignore the underlying characteristics of the organism, such as intelligence, if knowledge about these factors did not improve our predictions and control of behavior. However, as Burt discusses later, such knowledge of intelligence does improve our predictions, and this is especially true with school children.

The definition of intelligence is a central issue in the articles which follow, especially those by Burt and Guilford. Whether intelligence is a single general (*g*) factor or a group of factors is at issue between them. Burt shows that intelligence was conceived as a single unitary quality until the eighteenth century, when "faculty" psychology and phrenology made their appearances. This is also the common-sense view—when we say that someone is intelligent we usually mean that he is smart in almost everything he does. If we say he is dull, we think of his dullness as being a similarly pervasive quality. Guilford, using a statistical technique known as factor analysis, maintains that intelligence is a much more complex phenomenon—it may be composed of as many as one hundred and twenty factors, all of which can vary in closeness of relationship to one another. According to this view, Johnny is not uniformly smart in all things he seeks to do. He may, for example, do very well in "convergent thinking," where he is expected to come up with a single correct answer on a multiple choice test; but he may usually do more poorly in "divergent thinking," where it is expected that he will come up with creative and novel, and not necessarily incorrect, answers that the teacher never dreamed of. The definition of intelligence that is chosen also influences the interpretation of data on the development of intelligence. Does wisdom grow with the years? We know we lose youth, but do we gain intelligence with age? Although the picture once looked fairly dismal, Nancy Bayley, in the research report which follows, has brightened up the forecast for adulthood, but she uses a definition of intelligence that is more akin to Guilford's than Burt's.

In the study of intelligence, as well as of other human characteristics, the psychology of individual differences has an embarrassing way of unearthing facts that do not conveniently fit current social and political ideologies. There have been instances where psycholo-

gists have risked their professional reputations because they forced such data to fit a congenial ideology. Given the prevailing democratic and liberal beliefs of Americans, it is easier to accept psychological research when it tells us that we frequently confuse differences in degree with differences in type. We are reminded that no matter what our religion, race, class, or nationality, we all share in the underlying quality being measured. It is harder for us to accept the fact that although all measurable differences are quantitative we experience great differences in behavior as qualitative. To most of us, it makes a difference whether it is a chilling winter day or a balmy summer day, even though the difference is only so many degrees Fahrenheit. Similar degrees of difference in the intelligence of our students makes a qualitative difference in their scholastic performance. Segregation has become a word with unpleasant connotations because of the current racial issue. Segregation of the bright from the not-so-bright has some of the same unpleasant connotations in the schools. The education student should contemplate how a democratic, public, compulsory education system can best solve this problem.

Relationship of Readings in Chapter 7

The article by Burt attempts to answer the question, "What is intelligence?" and in doing so, traces the history of the concept from the introspection of Plato to the present-day statistical approach. Whereas Burt accepts the idea of a *g*-factor, or single general intelligence, Guilford is a leading representative of the group-factors school. The latter discusses how the various factors are tested and the import this theory has for education. Bayley is concerned with whether or not intelligence grows, and how much, and when. Her definition of intelligence seems closer to Guilford's, but some *g*-factor seems to remain.

SIR CYRIL BURT

Editor, *British Journal of Statistical Psychology*

The Meaning and Assessment of Intelligence*

American education is based on belief in individual development and social improvement. We like to believe that in short spaces of time we can radically alter almost any state of human affairs. We are a nation of fantastic builders. In a very short period of time we have built a great number of schools, and we have plans for many more. Sometimes, however, we seem to confuse the body of the school building with the human resources, both teachers and students, who also, and more significantly, affect the educational change we hope to bring about. The same type of physical assault on the environment, which succeeds in producing the school building, often fails when we are confronted with limitations in human potential.

If intelligence is more a product of our genes than of our environment, we are confronted with very real limitations in human potential. Our democratic philosophy, combined with a pervasive sense of mastery of our physical environment, makes this a peculiarly difficult fact to face: intelligence has less to do with the environment and the future than with our ancestry and the past. This in itself may make a biological view of intelligence easier to accept in England than in America. The different points of view are exemplified by Skinner (American), who believes that adoption of modern techniques of education and the application of reinforcement theory open limitless educational vistas, and Burt (English), who seems to accept an intellectual élite and a wide range of individual differences as biological facts which no society or educational system can ignore.

* Reprinted and abridged with the permission of the author and The Eugenics Society from article of same title, *The Eugenics Review*, 47 (1955), 81-91.

Burt defines intelligence as a general unitary concept, a definition that contrasts sharply with the definition given by Guilford, who defines it in terms of group factors in the selection following Burt's. Burt marshalls a great deal of evidence to support his position: an introspectionist philosophical tradition beginning with Plato, Aristotle, and Cicero and continuing to the eighteenth century; the study of intelligence in modern biology, physiology, and experimental psychology; and the study of individual differences combined with modern statistics. In examining this evidence and the arguments Burt uses to support his position, the student may consider the following questions: (1) What are the major aspects of Burt's concept of intelligence? (2) What additional evidence is needed to support his concept? (3) Why does he reject the group-factor concept of intelligence? (4) What conclusions can be drawn from the study in which he compared the intelligence of brothers and sisters? (5) What implications do his conclusions have for classroom practice?

A mere glance at the relevant literature will quickly show that intelligence is not a conception "introduced by a small group of statistical psychologists." Nor is the term itself "a word of popular speech" whose meaning has recently been restricted and distorted by psychological specialists. It is, and always has been, a technical term introduced to designate a technical concept. And the concept itself has been reached and clarified by inquirers working along half a dozen different lines. Observational psychology, introspective psychology, experimental psychology, the speculations of the biologist, the theories of the neurologist, and finally the objective study of individual differences, each has contributed valuable evidence. The application of statistical methods has come only at the very end; their function has mainly been to decide between alternative explanations of certain observable facts, and so to clinch and confirm what had been provisionally inferred on far more concrete grounds. May I therefore begin by briefly tracing the history of the concept?

I. OBSERVATION AND INTROSPECTION

The basic notion goes back to the days when the human mind first became the subject of philosophic curiosity. Plato, a shrewd observer of individuals, was, as Galton has so often reminded us, the first to recognize the social implications of mental heredity and to advocate something very like a eugenic policy. His psychological disquisitions are incidental and sporadic; but they had a profound influence on later thought.

He draws a clear contrast between "nature" and "nurture" (φύσις and τροφή); and he then goes on to distinguish three "parts" of the soul—the "rational" or intellectual (τὸ λογιστικόν) having its seat in the brain, and "appetite" (ἐπιθυμία) and "spirit" (θυμός) located respectively in the belly and chest. This threefold distinction has often been compared with the modern distinction between "cognition," "affection," and "conation"—the intellectual, emotional, and moral elements in human behaviour. But none of these modern terms accurately expresses what Plato was trying to convey. In a famous passage (*Phaedrus*, 253d) he uses an analogy which gives a better notion of the difference: the first element he compares to the charioteer who holds the reins, and the others to a pair of horses who draw it: the former guides, the latter provide the power; the former is the *cybernetic* element, the latter the *dynamic*.

And, says Plato, since men differ so widely in their innate characteristics, they should, from childhood upwards, be subjected to tests, so that each can be educated, and eventually employed, as his native gifts require. The rulers are to be men pre-eminent for their intellectual capacity or "wisdom"—"men of gold rather than of silver, iron, or brass."

Thus, for Plato the natural inequality of man is itself one of the most profound and ill-recognized of all political problems. It threatened the democracy of Athens, and it threatens the democratic state today: "It is the source at once of the injustice that we must seek to correct and of the justice or civic harmony that will enable us to correct it." Plato would have his citizens believe, "as though an oracle had foretold it, that the city will perish when men of iron or brass take over its control": or, as he puts it elsewhere, "the ship of state is bound to founder if the unruly crew, whose

job it is to manage the sails, but who are in no way 'cybernetic' (i.e. good at finding and steering a course), selfishly seize the helm."

Aristotle's discussion is more methodical, and issues in a more systematic classification. Here for the first time we meet a clear distinction between actual process and mere capacity or "power" (*δύναμις*). While lecturing I am actually talking; when asleep, I have the power to talk; when newborn, "with no language but a cry," I have the power to acquire the power to talk. The distinction is of course applicable in non-psychological fields as well as psychological: as applied to the latter it is the basis of our concept of mental capacity.

In what is virtually the first textbook of psychology Aristotle substitutes a twofold classification for Plato's threefold; and his main contrast is drawn between what he calls the "dianoetic" (cognitive or intellectual) capacities of the mind and the "orectic" (emotional and moral). The cognitive capacities manifest themselves at four successive levels—sensation, imagination, memory, and reasoning. . . . There is, however, no sharp separation between the various levels or the different parts or faculties. "Soul in fact is homœomerous, like a tissue" (i.e. it is not a collection of distinct organs): "with Aristotle sensation is regarded as itself a discriminative capacity from which the higher acts of cognition are reached by a continuous development."

Throughout, it will be noted, Aristotle formulates his classification of mental activities in terms of conscious contents, and so gives it an introspective rather than a behaviouristic character—a bias which has only recently been corrected.

Here then we have the origin of both the concept and the term. From Aristotle and Cicero they descended to the mediæval schoolmen; and the scholastic theories in turn developed into the cut-and-dried schemes of the faculty psychologists and their phrenological followers. All of them continued to contrast intellectual capacities, which they termed abilities or "faculties," with emotional or moral capacities, which they termed "propensities"; but none recognized any "general" ability over and above the more specific faculties. And according to the phrenologists each distinguishable mental function was due to the activity of a separate "organ" or "centre" in the brain. The whole picture is one that Plato would

instantly have repudiated, since he himself ridicules those who thought of the mind as a sort of "Trojan horse," containing within itself a collection of active homunculi, each with its own special task. Although the later psychologists of the nineteenth century, including both associationists and their critics, were united in rejecting it, the traditional theory of faculties continued to enjoy a considerable vogue among medical and educational writers. To this day, indeed, teachers, educational officials, school medical officers and psychiatrists constantly drop into the vocabulary of the faculty school when they attempt a character-sketch of any child or patient; and contemporary critics of the concept of "intelligence" regularly assume that its sponsors intend it as yet another "faculty" in the sense defined by the Scottish philosophers and their physiological interpreters.

2. BIOLOGICAL

In this country the conversion of psychology from a branch of philosophy into a branch of natural science was the work not of the physiologists but of the biologists, particularly the leaders of the evolutionary school—Spencer, Darwin and their disciples. Spencer, following Aristotle and the Thomists rather than Plato and Kant, recognized only two main aspects of mental life—the cognitive and the affective. All cognition (he explains) involves both an analytic or discriminative and a synthetic or integrative process; and its essential function is to enable the organism to adjust itself more effectively to a complex and ever-changing environment. During the evolution of the animal kingdom, and during the growth of the individual child (which, he assumes, briefly recapitulates the evolution of the race), the fundamental capacity of cognition becomes progressively more and more specialized and more and more comprehensive, and so differentiates into a hierarchy of cognitive abilities—sensory, perceptual, associative, and relational, much as the trunk of a tree sprouts into boughs, branches, and twigs. To designate the basic quality common to all these more specific forms he adopts the term "intelligence."

Spencer's evolutionary theories were at first taken up with keener enthusiasm on the Continent than in this country. Taine, the leader

of the new empirical school in France, expounded them in his monograph *De l'intelligence* (1870); Ribot amplified them still further in *L'hérédité psychologique* (1873); and their version provided the starting point for the work of their more celebrated disciple, Alfred Binet (*L'étude expérimentale de l'intelligence*, 1903, and later papers). In Switzerland Spencer's views inspired the genetic studies of Claparède and of his pupil, Jean Piaget.¹ Both these adopt a standpoint that is frankly biological. Piaget, in language reminiscent of Plato, contrasts the "directive" and "dynamic" elements in mental life: "every action," he says, "involves an energetic or affective aspect, and a structural, regulative, or cognitive aspect. . . . Intelligence is not a faculty: it is the generic term indicating the organism's relative efficiency in organizing or structuring mental activity in order to adjust itself to changing circumstances." And he propounds, as a result of first-hand observations of the developing child, a hierarchical theory of "levels," less schematic and more exact, yet on the whole strikingly similar to that of Herbert Spencer.

3. PHYSIOLOGICAL

While in France and Britain scientific psychology was regarded as a branch of biology, in Germany it was treated as a branch of physiology. The earliest experiments on cerebral localization seemed to indicate something rather like a modified phrenological theory—the functions localized in the various cortical areas being of a somewhat simpler kind than the traditional faculties. Wundt quotes with approval Spencer's principle that mental organization merely reflects the underlying neurological organization, and consequently regards *Intelligenz* as a property of the central nervous system. There is, however, no localized "Organ der Intelligenz": *Intelligenz* is "simply a name for the varying degrees of efficiency in the fundamental cognitive process"—a process which he prefers to call "apperception"—i.e. "attention regarded as a process of synthesis." It operates on various levels; and he too gives a schematic diagram of the way the nervous system is organized, plainly suggested by Spencer's description.

¹ J. Piaget, *The Psychology of Intelligence* (1950), and *The Origin of Intelligence in the Child* (1953).

Wundt's scheme is avowedly hypothetical. But later studies of the structure and functions of the nervous system went far to confirm the general accuracy of these views. The clinical work of Hughlings Jackson and the experimental investigations of Sherrington lent strong support to the theory of a "neural hierarchy," with a definite order of evolution for the various levels. Within the adult brain there are marked differences in the architecture of different parts and of the different cell layers clearly discernible under the microscope; and these differences or specializations emerge progressively during the earliest months of infant life. At the same time, the examination of the cortex in mental defectives and in normal persons indicates that the quality of the nervous tissue in any given individual tends to be predominantly the same throughout. Defectives, for example, exhibit a "general cerebral immaturity," and their nerve-cells tend to be "visibly deficient in number, branching, and regularity of arrangement in every part of the cortex." After all, as Sherrington points out, much the same is true of almost every tissue of which the human frame is composed—of a man's skin, bones, hair, or muscles: each is of the same general character all over the body, although minor local variations are usually discernible.

4. INDIVIDUAL PSYCHOLOGY

Most of the writers I have so far mentioned were interested chiefly in problems of *general* psychology. The first to apply scientific methods to the study of *individual* psychology was Galton himself. Spencer had maintained that the basic characteristics of the human mind were innate—transmitted as part of the common racial endowment. Galton went farther and maintained that individual differences in these characteristics might also be inherited or at least inborn. When he first commenced his inquiries on mental inheritance, the prevailing hypothesis among those who attempted to describe individual differences was, as we have seen, that of the faculty school. Galton quickly became convinced that a theory of wholly specific faculties was of itself quite inadequate to account for the facts he had accumulated.

As a corrective, he introduced the distinction between what he termed "general ability" and "special aptitudes." He recognizes

three main sources of individual achievement—cognitive capacities (or “abilities”), emotional or affective characteristics (such as “interest” or “zeal”), and moral or conative characteristics (notably “a will to work”). He focuses attention mainly on the first, since “natural” ability must inevitably set a limit to what interest or industry, even in the most favourable circumstances, can possibly achieve. Most writers, he argues, “lay too much stress upon apparent specialities, thinking that, because a man is devoted to some particular pursuit, he could not have succeeded in anything else; they might as well say that, because a youth has fallen in love with a brunette, he could not possibly have fallen in love with a blonde. He may or may not have had more natural liking for the former type of beauty than for the latter; but it is as probable as not that the affair was mainly or wholly due to a *general* amorousness. It is just the same with intellectual pursuits.”

Galton does not deny the existence of special capacities. Indeed, he cites instances in which memory, literary ability, musical ability, and artistic talent, run through several members of the same family. In some cases the specialization may be due to family tradition or to home environment, though this could scarcely explain the “prodigies of memory”; but, in the main, he says, the pedigrees and case-studies given in his book demonstrate “in how small a degree intellectual eminence can be considered as due to purely special powers.” His data suggest that individual differences in “natural ability” are distributed in accordance with the normal curve, i.e. much like differences in other human characteristics which are mainly innate, such as bodily size or stature; and he prints a tabular classification of frequencies, which, he holds, “may apply to special just as truly as to general ability.”

Binet was greatly influenced by Galton’s theories. Like Galton he distinguishes between *acquired* knowledge or skill (to be assessed by a “pedagogical scale”) and *native* abilities (to be assessed by a “psychological scale”). Like Galton, too, he firmly believes in the notion of *general* ability, which he contrasts “with partial aptitudes.” To designate this native general ability, he prefers the simple Spenserian name, “intelligence.” He gives us a popular but fairly clear account of “the meaning to be given to that word, so wide and comprehensive, intelligence. . . . Nearly all the phenomena with which

psychology is concerned are phenomena of intelligence—sensation, perception, as much as reasoning. . . . And it would seem that in the phenomena of intelligence there is a fundamental faculty, deficiency in which is of the utmost importance for practical life: this faculty," he continues, "is variously described as common sense, judgment, the capacity of adjusting oneself to circumstances" (the last is Spencer's definition). Since it enters into every cognitive process, tests of any such process might in theory be used to assess it. But, he adds, "it is neither necessary nor possible to test *all* the child's psychological processes." There is "a hierarchy among the diverse manifestations of intelligence"; the more complex and more specialized mature at later stages in a progressive order that is relatively fixed. Hence the crucial test for an individual at any given stage of development will be the hardest cognitive processes of which he is capable.

Such views did not escape criticism. A hypothesis which postulated *both* a general ability *and* a number of "partial" or "special" aptitudes seemed to assume two types of capacity where one would suffice. Writers on applied psychology, including the compilers of the more popular educational and psychiatric textbooks, usually rejected the notion of a central cognitive activity as a needless philosophical abstraction, and contended that a collection of special abilities or faculties accorded best with their practical experience. On the other hand, most of the writers on pure psychology treated the doctrine of special faculties as obsolete. There was, they held, and there could be, only one form of cognitive activity, though they failed to agree about its actual nature. The older associationists, such as Mill and his followers, maintained that it was "association"—the "process by which we learn"; the younger members of the school, like Bain and Sully, maintained that it was "sensory discrimination." Of their various opponents, both the neo-Kantian philosophers like Ward and the Herbartian psychologists like Stout and Adams, argued that it was "apperception" or "attention": "when we feel, perceive, or remember a thing [says Ward] common sense thinks the object is the same, while the mental faculty differs; actually there is only a single subjective activity—attention, and what we attend to are different presentations of the object." Finally, several of Ward's disciples, like Maxwell Garnett, insisted that at-

- (v) The statistical "general factor," i.e., "the factor common to all tests of a given battery," g_s , say; it is this more abstract quantity with which Thomson and Thurstone are concerned in their controversies with Spearman.
- (vi) "Intelligence as the layman understands the word," g_t .³

Which are we to choose? Most investigators simply compare an initial set of test-measurements with a later set, each derived from written tests applied on a single occasion only, i.e. g_t . But the correlation between two such tests, even applied with only a minimal time interval, is still never more than about 0.85 or 0.90. Hence, much of the imperfection shown by such correlations must be due to defects in the methods of assessment; they throw no light (as is so often alleged) on the supposed instability of γ . After all, no experienced psychologist would diagnose a child as feeble-minded on such a basis. He would invariably check the crude test-results by the child's case-history and the teacher's report, and in case of doubt retest him on a different day with one or more individual tests. Hence in what follows I shall be concerned chiefly with measurements of the third kind, g_s , i.e. assessments checked and corrected in this way. With these adjusted measurements the correlations are appreciably higher than those commonly reported for the unadjusted g_t .

For the cases I have been able to follow up, the correlations during the school period diminish progressively from 0.93 after one year to 0.74 after six years. The correlations with assessments secured in early adult life (i.e. after ten or fifteen years) average 0.61; and with assessments for the children of the original testees they average 0.32.

Correlations between parents and children are apt to vary somewhat erratically, in part no doubt because assessments of adult intelligence are bound to be more or less inaccurate. Let us therefore compare measurements for brothers and sisters who are all of school age. Miss Howard and I took a batch of 268 ten-year-olds each of whom had at least one sib attending school (i.e. aged eight to twelve) and who were so chosen as to be fairly representative of the total school population (excluding pathological defectives). They were divided into four equal groups: (i) bright, (iia) bright average, (iib)

³ I take this definition from R. B. Cattell, *Factor Analysis* (1952), p. 424.

dull average, (iii) dull; and the middle groups were pooled. When any child had more than one sib, the sibs' assessments were averaged. The bivariate frequency-distribution so obtained, expressed in the form of percentages, is shown in the table below.

Table 1

FREQUENCY DISTRIBUTION OF BRIGHT, AVERAGE, AND DULL SIBS

Selected Children	<i>Bright</i>	Sibs <i>Average</i>	<i>Dull</i>	Total
Bright	12.8	10.0	2.2	25.0
Average	11.2	29.6	9.2	50.0
Dull	1.3	8.9	14.8	25.0
Total	25.3	48.5	26.2	100.0

On Mendelian principles, it is easy to show that, assuming variations in intelligence are produced by a large number of genes, then the expected proportions in the several rows would be 4:4:1, 3:8:3, and 1:4:4 respectively; i.e., with subgroups of 25 and 50 we should expect the figures to read 11.1, 11.1, 2.8; 10.7, 28.6, 10.7; and 2.8, 11.1, 11.1. In the middle row the observed figures conform quite closely with expectation. The excess of bright sibs in the top row and of dull sibs in the bottom row is probably due to the fact that like tends to marry like. Assortative mating would obviously raise the apparent correlation. On the other hand, the inaccuracies in the assessments would tend to lower it. Allowing for such minor disturbances, the frequencies clearly suggest that we are dealing with a trait that is, in the main, the effect of multi-factor or "polygenic" inheritance. One special merit of the Mendelian theory is that it reminds us that genetic conditions are responsible not merely for resemblances between members of the same family but also for differences. It explains what on any other theory remains such an unaccountable paradox, namely, the occasional occurrence of exceptional bright children in homes where the dullness of the parents and the handicaps of the environment would, one might have supposed, condemned the offspring to hopeless failure.

Figures like the foregoing cannot of themselves provide conclusive proof that the characteristic we are seeking to assess is an innate

and therefore permanent characteristic. But they plainly offer strong corroboration for what is, on antecedent grounds, a highly plausible hypothesis. As with most generalizations in the field of individual psychology, our acceptance of such a conclusion must rest, not on any one decisive inquiry, but on inferences reached by half a dozen different lines of approach and set forth in numerous independent researches. The evidence so far available I have summarized in some detail in other publications, and accordingly I need not repeat it here. Roughly speaking, an impartial analysis would seem to indicate that very nearly 90 per cent of the variance exhibited by assessments for a complete age-group is attributable to the genetic constitution of the various individuals and that approximately half of this (i.e. 45 per cent of the variance) is attributable to what is loosely called heredity (i.e. predictable from characteristics of near relatives).

Finally, what is the actual distribution of "intelligence" as we have defined it? Like variations in stature, variations in intelligence (γ), as Galton himself believed, follow to a close approximation the normal curve. But, what is much more important, not only is there "a continuity of natural ability," but "the range of mental power between the greatest and least of English intellect is enormous." Surveys carried out in London and elsewhere show that, if we take a random sample of 1,000 children aged ten by the calendar, and exclude all pathological cases, the dullest will have a mental age of only five, the brightest a mental age of approximately fifteen, and between these two extremes every intermediate grade will be found. There is a larger proportion of bright children in the upper classes and a smaller proportion in the lower, but the several classes exhibit a wide overlapping. Moreover, were we to divide the total population into the non-professional and professional classes, then, simply because the former are far more numerous, I calculate that, in the former, we should find approximately three times as many "very bright" children (say, sufficiently able to pass an honours examination) as in the latter.

If the views that I have put forward are correct, it is clear that these inequalities in native ability, as Plato long ago foresaw, present the democratic state with profound and far-reaching problems—problems which even today are scarcely recognized and which

have been attacked only in the most tentative fashion. So far as the child is concerned, it is plainly imperative that the education authority should seek to determine as accurately as possible the natural potentialities of each one, and, having done so, provide him with the education best suited to his needs, and finally, before it leaves him, help to select that kind of vocation for which his gifts may seem to have marked him out. In this way, and in this way alone, can we hope to realize "that ideal polity in which the apparent injustices of nature are reconciled and harmonized by the wisdom and justice of man."

J. P. GUILFORD

University of Southern California

Three Faces of Intellect*

In his course in educational psychology the student very likely will meet the terms "faculty psychology" and "mental discipline school." Burt has already referred to the former (see pp. 422-423). Faculty psychology was a theory of how the mind functions, and mental discipline was the method of education or training purportedly based on it. The various faculties of the mind were metaphorically described as muscles, so that one had a "memory muscle," a "reasoning muscle," a "judgment muscle," etc. The chief function of the teacher was to provide the student with materials which would exercise these mental muscles. These materials might consist of Latin and Greek texts, mathematics, logic, etc. These seemed "harder" subjects and therefore better muscle builders. The theory allowed teachers to ignore the specific ultimate uses education would be put to because it took (without testing) a blindly optimistic view of transfer of training (see pp. 32-33). It cavalierly assumed

* Reprinted and abridged with the permission of the author and the American Psychological Association from the article of the same title, *American Psychologist*, 14 (1959), 469-479.

that proficiency in Ciceronian Latin and Aristotelian logic could be used equally well by doctors, lawyers, bankers, and certified public accountants. After all, it was contended, what students really learned was *how* to think; their minds became finely polished instruments in the mental discipline school.

The recent shift in our schools to intellectual interests has accompanied a revival of the mental discipline school. In this article, for example, Guilford describes "basic" mental processes (or operations) such as "cognition," "memory," etc. He argues for the training of the mind and for less emphasis on the teaching of "skills and habits." His position, therefore, seems much closer to that of Bruner, who is also interested in complex mental processes. Bruner's "structure of knowledge" is comparable to Guilford's "structure of intellect." Bruner has described the "act of discovery"; Guilford has defined cognition as "discovery," and he goes on to discuss "divergent thinking" as a kind of cognitive flexibility resulting in a variety of novel responses (see pp. 444-445). Skinner and Kendler, on the other hand, desire to introduce much more specificity in the learning situation. Both urge the teacher to know the exact nature of the response he wants the student to make and to arrange the stimulus situation, or the "contingencies of reinforcement," in such a way as to guarantee the correct response. Kendler refers to knowing more about the mediating events in higher forms of thinking (p. 391) and Skinner sees the need to analyze thinking into its specific component behaviors (p. 178) in order to teach it more effectively.

In analyzing the article by Guilford, the student may find the following questions helpful: (1) What evidence is presented to show that intelligence is a group of factors rather than a single general factor? How would Burt evaluate this evidence? (2) Why did Guilford find it necessary to give examples of many test questions? How do these affect his definition of intelligence? (3) In what ways does his model of intelligence include emotional as well as cognitive behavior? (4) In what ways could the teacher promote "discovery" and "divergent thinking"? Do your answers conform to the views of Bruner and Kersh (pp. 254-287)?

Our knowledge of the components of human intelligence has come about mostly within the last 25 years. The major sources of this information in this country have been L. L. Thurstone and his associates, the wartime research of psychologists in the United States Air Forces, and more recently the Aptitudes Project at the University of Southern California, now in its tenth year of research on cognitive and thinking abilities. The results from the Aptitudes Project that have gained perhaps the most attention have pertained to creative-thinking abilities. These are mostly novel findings. But to me, the most significant outcome has been the development of a unified theory of human intellect, which organizes the known, unique or primary intellectual abilities into a single system called the "structure of intellect." It is to this system that I shall devote the major part of my remarks, with very brief mentions of some of the implications for the psychology of thinking and problem solving, for vocational testing, and for education.

The discovery of the components of intelligence has been by means of the experimental application of the method of factor analysis. It is not necessary for you to know anything about the theory or method of factor analysis in order to follow the discussion of the components. I should like to say, however, that factor analysis has no connection with or resemblance to psychoanalysis. A positive statement would be more helpful, so I will say that each intellectual component or factor is a unique ability that is needed to do well in a certain class of tasks or tests. As a general principle we find that certain individuals do well in the tests of a certain class, but they may do poorly in the tests of another class. We conclude that a factor has certain properties from the features that the tests of a class have in common. I shall give you very soon a number of examples of tests, each representing a factor.

The Structure of Intellect

Although each factor is sufficiently distinct to be detected by factor analysis, in very recent years it has become apparent that the factors themselves can be classified because they resemble one an-

other in certain ways. One basis of classification is according to the basic kind of process or operation performed. This kind of classification gives us five major groups of intellectual abilities: factors of cognition, memory, convergent thinking, divergent thinking, and evaluation.

Cognition means discovery or rediscovery or recognition. Memory means retention of what is cognized. Two kinds of productive-thinking operations generate new information from known information and remembered information. In divergent-thinking operations we think in different directions, sometimes searching, sometimes seeking variety. In convergent thinking the information leads to one right answer or to a recognized best or conventional answer. In evaluation we reach decisions as to goodness, correctness, suitability, or adequacy of what we know, what we remember, and what we produce in productive thinking.

A second way of classifying the intellectual factors is according to the kind of material or content involved. The factors known thus far involve three kinds of material or content: the content may be figural, symbolic, or semantic. Figural content is concrete material such as is perceived through the senses. It does not represent anything except itself. Visual material has properties such as size, form, color, location, or texture. Things we hear or feel provide other examples of figural material. Symbolic content is composed of letters, digits, and other conventional signs, usually organized in general systems, such as the alphabet or the number system. Semantic content is in the form of verbal meanings or ideas, for which no examples are necessary.

When a certain operation is applied to a certain kind of content, as many as six general kinds of products may be involved. There is enough evidence available to suggest that, regardless of the combinations of operations and content, the same six kinds of products may be found associated. The six kinds of products are: units, classes, relations, systems, transformations, and implications. So far as we have determined from factor analysis, these are the only fundamental kinds of products that we can know. As such, they may serve as basic classes into which one might fit all kinds of information psychologically.

The three kinds of classifications of the factors of intellect can be

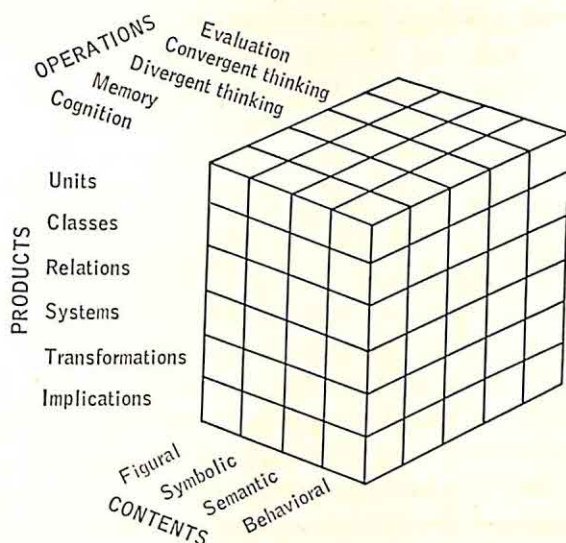


Fig. 1. A Cubical Model Representing the Structure of Intellect.

represented by means of a single solid model, shown in Figure 1. In this model, which we call the "structure of intellect," each dimension represents one of the modes of variation of the factors. Along one dimension are found the various kinds of operations, along a second one are the various kinds of products, and along the third are various kinds of content. Along the dimension of content a fourth category has been added, its kind of content being designated as "behavioral." This category has been added on a purely theoretical basis to represent the general area sometimes called "social intelligence." More will be said about this section of the model later.

In order to provide a better basis for understanding the model and a better basis for accepting it as a picture of human intellect, I shall do some exploring of it with you systematically, giving some examples of tests. Each cell in the model calls for a certain kind of ability that can be described in terms of operation, content, and product, for each cell is at the intersection of a unique combination of kinds of operation, content, and product. A test for that ability would have the same three properties. In our exploration of the model, we shall take one vertical layer at a time, beginning with the

front face. The first layer provides us with a matrix of 18 cells (if we ignore the behavioral column for which there are as yet no known factors) each of which should contain a cognitive ability.

THE COGNITIVE ABILITIES

We know at present the unique abilities that fit logically into 15 of the 18 cells for cognitive abilities. Each row presents a triad of similar abilities, having a single kind of product in common. The factors of the first row are concerned with the knowing of units. A good test of the ability to cognize figural units is the Street Gestalt Completion Test. In this test, the recognition of familiar pictured objects in silhouette form is made difficult for testing purposes by blocking out parts of those objects. There is another factor that is known to involve the perception of auditory figures—in the form of melodies, rhythms, and speech sounds—and still another factor involving kinesthetic forms. The presence of three factors in one cell (they are conceivably distinct abilities, although this has not been tested) suggests that more generally, in the figural column, at least, we should expect to find more than one ability. A fourth dimension pertaining to variations in sense modality may thus apply in connection with figural content. The model could be extended in this manner if the facts call for such an extension.

The ability to cognize symbolic units is measured by tests like the following:

Put vowels in the following blanks to make real words:

P _ W _ R
M _ RV _ L
C _ RT _ N

Rearrange the letters to make real words:

R A C I H
T V O E S
K L C C O

The first of these two tests is called Disemvoweled Words, and the second Scrambled Words.

The ability to cognize semantic units is the wellknown factor of verbal comprehension, which is best measured by means of a vocabulary test, with items such as:

GRAVITY means _____
 CIRCUS means _____
 VIRTUE means _____

From the comparison of these two factors it is obvious that recognizing familiar words as letter structures and knowing what words mean depend upon quite different abilities.

For testing the abilities to know classes of units, we may present the following kinds of items, one with symbolic content and one with semantic content.

Which letter group does not belong?

XECM PVAA QXIN VTRO

Which object does not belong?

clam tree oven rose

A figural test is constructed in a completely parallel form, presenting in each item four figures, three of which have a property in common and the fourth lacking that property.

The three abilities to see relationships are also readily measured by a common kind of test, differing only in terms of content. The well-known analogies test is applicable, two items in symbolic and semantic form being:

JIRE : KIRE :: FORA : KORE KORA LIRE GORA GIRE
 poetry : prose :: dance : music walk sing talk jump

Such tests usually involve more than the ability to cognize relations, but we are not concerned with this problem at this point.

The three factors for cognizing systems do not at present appear in tests so closely resembling one another as in the case of the examples just given. There is nevertheless an underlying common core of logical similarity. Ordinary space tests, such as Thurstone's Flags, Figures, and Cards or Part V (Spatial Orientation) of the Guilford-Zimmerman Aptitude Survey (GZAS), serve in the figural column. The system involved is an order or arrangement of objects in space. A system that uses symbolic elements is illustrated by the Letter Triangle Test, a sample item of which is:

```

      —
    d   —
  b     e   —
a     c   f   ?

```

What letter belongs at the place of the question mark?

The ability to understand a semantic system has been known for some time as the factor called general reasoning. One of its most faithful indicators is a test composed of arithmetic-reasoning items. That the phase of understanding only is important for measuring this ability is shown by the fact that such a test works even if the examinee is not asked to give a complete solution; he need only show that he structures the problem properly. For example, an item from the test Necessary Arithmetical Operations simply asks what operations are needed to solve the problem:

A city lot 48 feet wide and 149 feet deep costs \$79,432. What is the cost per square foot?

- A. add and multiply
- B. multiply and divide
- C. subtract and divide
- D. add and subtract
- E. divide and add

Placing the factor of general reasoning in this cell of the structure of intellect gives us some new conceptions of its nature. It should be a broad ability to grasp all kinds of systems that are conceived in terms of verbal concepts, not restricted to the understanding of problems of an arithmetical type.

Transformations are changes of various kinds, including modifications in arrangement, organization, or meaning. In the figural column for the transformations row, we find the factor known as visualization. Common measuring instruments for this factor are the surface-development tests, and an example of a different kind is Part VI (Spatial Visualization) of the GZAS. A test of the ability to make transformations of meaning, for the factor in the semantic column, is called Similarities. The examinee is asked to state several ways in which two objects, such as an apple and an orange, are alike. Only by shifting the meanings of both is the examinee able to give many responses to such an item.

In the set of abilities having to do with the cognition of implications, we find that the individual goes beyond the information given, but not to the extent of what might be called drawing conclusions. We may say that he extrapolates. From the given information he expects or foresees certain consequences, for example. The two factors found in this row of the cognition matrix were first called "foresight" factors. Foresight in connection with figural material can be tested by means of paper-and-pencil mazes. Foresight in connection

with ideas, those pertaining to events, for example, is indicated by a test such as Pertinent Questions:

In planning to open a new hamburger stand in a certain community, what four questions should be considered in deciding upon its location?

The more questions the examinee asks in response to a list of such problems, the more he evidently foresees contingencies.

THE MEMORY ABILITIES

The area of memory abilities has been explored less than some of the other areas of operation, and only seven of the potential cells of the memory matrix have known factors in them. These cells are restricted to three rows: for units, relations, and systems. The first cell in the memory matrix is now occupied by two factors, parallel to two in the corresponding cognition matrix: visual memory and auditory memory. Memory for series of letters or numbers, as in memory span tests, conforms to the conception of memory for symbolic units. Memory for the ideas in a paragraph conforms to the conception of memory for semantic units.

The formation of associations between units, such as visual forms, syllables, and meaningful words, as in the method of paired associates, would seem to represent three abilities to remember relationships involving three kinds of content. We know of two such abilities, for the symbolic and semantic columns. The memory for known systems is represented by two abilities very recently discovered (Christal, 1958).^{*} Remembering the arrangement of objects in space is the nature of an ability in the figural column, and remembering a sequence of events is the nature of a corresponding ability in the semantic column. The differentiation between these two abilities implies that a person may be able to say where he saw an object on a page, but he might not be able to say on which of several pages he saw it after leafing through several pages that included the right one. Considering the blank rows in the memory matrix, we should expect to find abilities also to remember classes, transformations, and implications, as well as units, relations, and systems.

^{*} R. E. Christal, "Factor analytic studies of visual memory," *Psychological Monographs*, 1958, 72, No. 13 (Whole No. 466).

THE DIVERGENT-THINKING ABILITIES

The unique feature of divergent production is that a *variety* of responses is produced. The product is not completely determined by the given information. This is not to say that divergent thinking does not come into play in the total process of reaching a unique conclusion, for it comes into play wherever there is trial-and-error thinking.

The well-known ability of word fluency is tested by asking the examinee to list words satisfying a specified letter requirement, such as words beginning with the letter "s" or words ending in "-tion." This ability is now regarded as a facility in divergent production of symbolic units. The parallel semantic ability has been known as ideational fluency. A typical test item calls for listing objects that are round and edible. Winston Churchill must have possessed this ability to a high degree. Clement Attlee is reported to have said about him recently that, no matter what problem came up, Churchill always seemed to have about ten ideas. The trouble was, Attlee continued, he did not know which was the good one. The last comment implies some weakness in one or more of the evaluative abilities.

The divergent production of class ideas is believed to be the unique feature of a factor called "spontaneous flexibility." A typical test instructs the examinee to list all the uses he can think of for a common brick, and he is given eight minutes. If his responses are: build a house, build a barn, build a garage, build a school, build a church, build a chimney, build a walk, and build a barbecue, he would earn a fairly high score for ideational fluency but a very low score for spontaneous flexibility, because all these uses fall into the same class. If another person said: make a door stop, make a paper weight, throw it at a dog, make a bookcase, drown a cat, drive a nail, make a red powder, and use for baseball bases, he would also receive a high score for flexibility. He has gone frequently from one class to another.

A current study of unknown but predicted divergent-production abilities includes testing whether there are also figural and symbolic abilities to produce multiple classes. An experimental figural test presents a number of figures that can be classified in groups of three in various ways, each figure being usable in more than one class

An experimental symbolic test presents a few numbers that are also to be classified in multiple ways.

A unique ability involving relations is called "associational fluency." It calls for the production of a variety of things related in a specified way to a given thing. For example, the examinee is asked to list words meaning about the same as "good" or to list words meaning about the opposite of "hard." In these instances the response produced is to complete a relationship, and semantic content is involved. Some of our present experimental tests call for the production of varieties of relations, as such, and involve figural and symbolic content also. For example, given four small digits, in how many ways can they be related in order to produce a sum of eight?

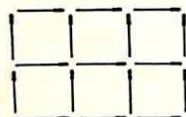
One factor pertaining to the production of systems is known as expressional fluency. The rapid formation of phrases or sentences is the essence of certain tests of this factor. For example, given the initial letters:

W_____ c_____ e_____ n_____

with different sentences to be produced, the examinee might write "We can eat nuts" or "Whence came Eve Newton?" In interpreting the factor, we regard the sentence as a symbolic system. By analogy, a figural system would be some kind of organization of lines and other elements, and a semantic system would be in the form of a verbally stated problem or perhaps something as complex as a theory.

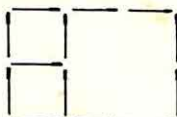
In the row of the divergent-production matrix devoted to transformations, we find some very interesting factors. The one called "adaptive flexibility" is now recognized as belonging in the figural column. A faithful test of it has been Match Problems. This is based upon the common game that uses squares, the sides of which are formed by match sticks. The examinee is told to take away a given number of matches to leave a stated number of squares with nothing left over. Nothing is said about the sizes of the squares to be left. If the examinee imposes upon himself the restriction that the squares that he leaves must be of the same size, he will fail in his attempts to do items like that in Figure 2. Other odd kinds of solutions are introduced in other items, such as overlapping squares and squares within squares, and so on. In another variation of

Item from the test match problems



A

Take away four matches in A, leaving three squares and nothing more. Answer: B.



B

Fig. 2. A Sample Item from the Test Match Problems. The Problem in This Item Is to Take Away Four Matches and Leave Three Squares. The Solution Is Given.

Match Problems the examinee is told to produce two or more solutions for each problem.

A factor that has been called "originality" is now recognized as adaptive flexibility with semantic material, where there must be a shifting of meanings. The examinee must produce the shifts or changes in meaning and so come up with novel, unusual, clever, or farfetched ideas. The Plot Titles Test presents a short story, the examinee being told to list as many appropriate titles as he can to head the story. One story is about a missionary who has been captured by cannibals in Africa. He is in the pot and about to be boiled when a princess of the tribe obtains a promise for his release if he will become her mate. He refuses and is boiled to death.

In scoring the test, we separate the responses into two categories, clever and nonclever. Examples of nonclever responses are: African Death, Defeat of a Princess, Eaten by Savages, The Princess, The African Missionary, In Darkest Africa, and Boiled by Savages. These titles are appropriate but commonplace. The number of such responses serves as a score for ideational fluency. Examples of clever responses are: Pot's Plot, Potluck Dinner, Stewed Parson, Gail or Boil, A Mate Worse Than Death, He Left a Dish for a Pot, Chaste in Haste, and A Hot Price for Freedom. The number of clever responses given by an examinee is his score for originality, or the divergent production of semantic transformations.

Another test of originality presents a very novel task so that any acceptable response is unusual for the individual. In the Symbol Production Test the examinee is to produce a simple symbol to stand for a noun or a verb in each short sentence, in other words to invent something like pictographic symbols. Still another test of originality asks for writing the "punch lines" for cartoons, a task that almost automatically challenges the examinee to be clever.

Thus, quite a variety of tests offer approaches to the measurement of originality, including one or two others that I have not mentioned.

Abilities to produce a variety of implications are assessed by tests calling for elaboration of given information. A figural test of this type provides the examinee with a line or two, to which he is to add other lines to produce an object. The more lines he adds, the greater his score. A semantic test gives the examinee the outlines of a plan to which he is to respond by stating all the details he can think of to make the plan work. A new test we are trying out in the symbolic area presents two simple equations such as $B - C = D$ and $z = A + D$. The examinee is to make as many other equations as he can from this information.

THE CONVERGENT-PRODUCTION ABILITIES

Of the 18 convergent-production abilities expected in the three content columns, 12 are now recognized. In the first row, pertaining to units, we have an ability to name figural properties (forms or colors) and an ability to name abstractions (classes, relations, and so on). It may be that the ability in common to the speed of naming forms and the speed of naming colors is not appropriately placed in the convergent-thinking matrix. One might expect that the thing to be produced in a test of the convergent production of figural units would be in the form of figures rather than words. A better test of such an ability might somehow specify the need for one particular object, the examinee to furnish the object.

A test for the convergent production of classes (Word Grouping) presents a list of 12 words that are to be classified in four, and only four, meaningful groups, no word to appear in more than one group. A parallel test (Figure Concepts Test) presents 20 pictured real objects that are to be grouped in meaningful classes of two or more each.

Convergent production having to do with relationships is represented by three known factors, all involving the "education of correlates," as Spearman called it. The given information includes one unit and a stated relation, the examinee to supply the other unit. Analogies tests that call for completion rather than a choice be-

tween alternative answers emphasize this kind of ability. With symbolic content such an item might read:

pots stop bard drab rats ?

A semantic item that measures education of correlates is:

The absence of sound is _____

Incidentally, the latter item is from a vocabulary-completion test, and its relation to the factor of ability to produce correlates indicates how, by change of form, a vocabulary test may indicate an ability other than that for which vocabulary tests are usually intended, namely, the factor of verbal comprehension.

Only one factor for convergent production of systems is known, and it is in the semantic column. It is measured by a class of tests that may be called ordering tests. The examinee may be presented with a number of events that ordinarily have a best or most logical order, the events being presented in scrambled order. The presentation may be pictorial, as in the Picture Arrangement Test, or verbal. The pictures may be taken from a cartoon strip. The verbally presented events may be in the form of the various steps needed to plant a new lawn. There are undoubtedly other kinds of systems than temporal order that could be utilized for testing abilities in this row of the convergent-production matrix.

In the way of producing transformations of a unique variety, we have three recognized factors, known as redefinition abilities. In each case, redefinition involves the changing of functions or uses of parts of one unit and giving them new functions or uses in some new unit. For testing the ability of figural redefinition, a task based upon the Gottschaldt figures is suitable. Figure 3 shows the kind of item for such a test. In recognizing the simpler figure within the structure of a more complex figure, certain lines must take on new roles.

In terms of symbolic material, the following sample items will illustrate how groups of letters in given words must be readapted to use in other words. In the test *Camouflaged Words*, each sentence contains the name of a sport or game:

I did not know that he was ailing.
To beat the Hun, tin goes a long way.

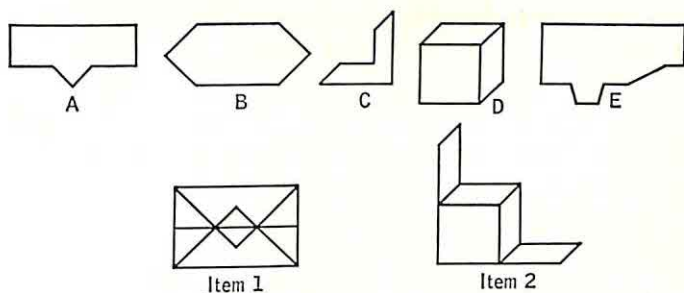


Fig. 3. Sample Items from a Test Hidden Figures, Based upon the Gottschaldt Figures. Which of the Simpler Figures Is Concealed within Each of the Two More Complex Figures?

For the factor of semantic redefinition, the Gestalt Transformation Test may be used. A sample item reads:

From which object could you most likely make a needle?

- A. a cabbage
- B. a splice
- C. a steak
- D. a paper box
- E. a fish

The convergent production of implications means the drawing of fully determined conclusions from given information. The well-known factor of numerical facility belongs in the symbolic column. For the parallel ability in the figural column, we have a test known as Form Reasoning, in which rigorously defined operations with figures are used. For the parallel ability in the semantic column, the factor sometimes called "deduction" probably qualifies. Items of the following type are sometimes used.

Charles is younger than Robert
 Charles is older than Frank
 Who is older: Robert or Frank?

EVALUATIVE ABILITIES

The evaluative area has had the least investigation of all the operational categories. In fact, only one systematic analytical study has

been devoted to this area. Only eight evaluative abilities are recognized as fitting into the evaluation matrix. But at least five rows have one or more factors each, and also three of the usual columns or content categories. In each case, evaluation involves reaching decisions as to the accuracy, goodness, suitability, or workability of information. In each row, for the particular kind of product of that row, some kind of criterion or standard of judgment is involved.

In the first row, for the evaluation of units, the important decision to be made pertains to the identity of a unit. Is this unit identical with that one? In the figural column we find the factor long known as "perceptual speed." Tests of this factor invariably call for decisions of identity, for example, Part IV (Perceptual Speed) of the GZAS or Thurstone's Identical Forms. I think it has been generally wrongly thought that the ability involved is that of cognition of visual forms. But we have seen that another factor is a more suitable candidate for this definition and for being in the very first cell of the cognitive matrix. It is parallel to this evaluative ability but does not require the judgment of identity as one of its properties.

In the symbolic column is an ability to judge identity of symbolic units, in the form of series of letters or numbers or of names of individuals.

Are members of the following pairs identical or not:

825170493 _____ 825176493

dkeltvmpa _____ dkeltvmpa

C. S. Meyerson _____ C. E. Meyerson

Such items are common in tests of clerical aptitude.

There should be a parallel ability to decide whether two ideas are identical or different. Is the idea expressed in this sentence the same as the idea expressed in that one? Do these two proverbs express essentially the same idea? Such tests exist and will be used to test the hypothesis that such an ability can be demonstrated.

No evaluative abilities pertaining to classes have as yet been recognized. The abilities having to do with evaluation where relations are concerned must meet the criterion of logical consistency. Syllogistic-type tests involving letter symbols indicate a different ability than the same type of test involving verbal statements. In the figural column we might expect that tests incorporating geometric reason-

ing or proof would indicate a parallel ability to sense the soundness of conclusions regarding figural relationships.

The evaluation of systems seems to be concerned with the internal consistency of those systems, so far as we can tell from the knowledge of one such factor. The factor has been called "experiential evaluation," and its representative test presents items like that in Figure 4 asking "What is wrong with this picture?" The things wrong are often internal inconsistencies.

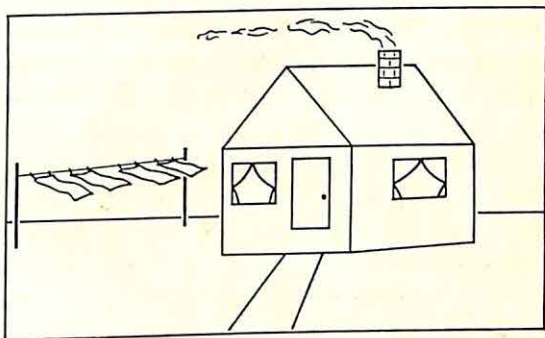


Fig. 4. A Sample Item from the Test Unusual Details. What Two Things Are Wrong with This Picture?

A semantic ability for evaluating transformations is thought to be that known for some time as "judgment." In typical judgment tests, the examinee is asked to tell which of five solutions to a practical problem is most adequate or wise. The solutions frequently involve improvisations, in other words, adaptations of familiar objects to unusual uses. In this way the items present redefinitions to be evaluated.

A factor known first as "sensitivity to problems" has become recognized as an evaluative ability having to do with implications. One test of the factor, the Apparatus Test, asks for two needed improvements with respect to each of several common devices, such as the telephone or the toaster. The Social Institutions Test, a measure of the same factor, asks what things are wrong with each of several institutions, such as tipping or national elections. We may say that defects or deficiencies are implications of an evaluative kind. Another interpretation would be that seeing defects and deficiencies are evaluations of implications to the effect that the various aspects of something are all right.

Some Implications of the Structure of Intellect

FOR PSYCHOLOGICAL THEORY

Although factor analysis as generally employed is best designed to investigate ways in which individuals differ from one another, in other words, to discover traits, the results also tell us much about how individuals are alike. Consequently, information regarding the factors and their interrelationships gives us understanding of functioning individuals. The five kinds of intellectual abilities in terms of operations may be said to represent five ways of functioning. The kinds of intellectual abilities distinguished according to varieties of test content and the kinds of abilities distinguished according to varieties of products suggest a classification of basic forms of information or knowledge. The kind of organism suggested by this way of looking at intellect is that of an agency for dealing with information of various kinds in various ways. The concepts provided by the distinctions among the intellectual abilities and by their classifications may be very useful in our future investigations of learning, memory, problem solving, invention, and decision making, by whatever method we choose to approach those problems.

FOR VOCATIONAL TESTING

With about 50 intellectual factors already known, we may say that there are at least 50 ways of being intelligent. It has been facetiously suggested that there seem to be a great many more ways of being stupid, unfortunately. The structure of intellect is a theoretical model that predicts as many as 120 distinct abilities, if every cell of the model contains a factor. Already we know that two cells contain two or more factors each, and there probably are actually other cells of this type. Since the model was first conceived, 12 factors predicted by it have found places in it. There is consequently hope of filling many of the other vacancies, and we may eventually end up with more than 120 abilities.

The major implication for the assessment of intelligence is that to know an individual's intellectual resources thoroughly we shall need

a surprisingly large number of scores. It is expected that many of the factors are intercorrelated, so there is some possibility that by appropriate sampling we shall be able to cover the important abilities with a more limited number of tests. At any rate, a multiple-score approach to the assessment of intelligence is definitely indicated in connection with future vocational operations.

Considering the kinds of abilities classified as to content, we may speak roughly of four kinds of intelligence. The abilities involving the use of figural information may be regarded as "concrete" intelligence. The people who depend most upon these abilities deal with concrete things and their properties. Among these people are mechanics, operators of machines, engineers (in some aspects of their work), artists, and musicians.

In the abilities pertaining to symbolic and semantic content, we have two kinds of "abstract" intelligence. Symbolic abilities should be important in learning to recognize words, to spell, and to operate with numbers. Language and mathematics should depend very much upon them, except that in mathematics some aspects, such as geometry, have strong figural involvement. Semantic intelligence is important for understanding things in terms of verbal concepts and hence is important in all courses where the learning of facts and ideas is essential.

In the hypothesized behavioral column of the structure of intellect, which may be roughly described as "social" intelligence, we have some of the most interesting possibilities. Understanding the behavior of others and of ourselves is largely nonverbal in character. The theory suggests as many as 30 abilities in this area, some having to do with understanding, some with productive thinking about behavior, and some with the evaluation of behavior. The theory also suggests that information regarding behavior is also in the form of the six kinds of products that apply elsewhere in the structure of intellect, including units, relations, systems, and so on. The abilities in the area of social intelligence, whatever they prove to be, will possess considerable importance in connection with all those individuals who deal most with other people: teachers, law officials, social workers, therapists, politicians, statesmen, and leaders of other kinds.

FOR EDUCATION

The implications for education are numerous, and I have time just to mention a very few. The most fundamental implication is that we might well undergo transformations with respect to our conception of the learner and of the process of learning. Under the prevailing conception, the learner is a kind of stimulus-response device, much on the order of a vending machine. You put in a coin, and something comes out. The machine learns what reaction to put out when a certain coin is put in. If, instead, we think of the learner as an agent for dealing with information, where information is defined very broadly, we have something more analogous to an electronic computer. We feed a computer information; it stores that information; it uses that information for generating new information, either by way of divergent or convergent thinking; and it evaluates its own results. Advantages that a human learner has over a computer include the step of seeking and discovering new information from sources outside itself and the step of programing itself. Perhaps even these steps will be added to computers, if this has not already been done in some cases.

At any rate, this conception of the learner leads us to the idea that learning is discovery of information, not merely the formation of associations, particularly associations in the form of stimulus-response connections. I am aware of the fact that my proposal is rank heresy. But if we are to make significant progress in our understanding of human learning and particularly our understanding of the so-called higher mental processes of thinking, problem solving, and creative thinking, some drastic modifications are due in our theory.

The idea that education is a matter of training the mind or of training the intellect has been rather unpopular, wherever the prevailing psychological doctrines have been followed. In theory, at least, the emphasis has been upon the learning of rather specific habits or skills. If we take our cue from factor theory, however, we recognize that most learning probably has both specific and general aspects or components. The general aspects may be along the lines of the factors of intellect. This is not to say that the individual's status in each factor is entirely determined by learning. We do not

know to what extent each factor is determined by heredity and to what extent by learning. The best position for educators to take is that possibly every intellectual factor can be developed in individuals at least to some extent by learning.

If education has the general objective of developing the intellects of students, it can be suggested that each intellectual factor provides a particular goal at which to aim. Defined by a certain combination of content, operation, and product, each goal ability then calls for certain kinds of practice in order to achieve improvement in it. This implies choice of curriculum and the choice or invention of teaching methods that will most likely accomplish the desired results.

Considering the very great variety of abilities revealed by the factorial exploration of intellect, we are in a better position to ask whether any general intellectual skills are now being neglected in education and whether appropriate balances are being observed. It is often observed these days that we have fallen down in the way of producing resourceful, creative graduates. How true this is, in comparison with other times, I do not know. Perhaps the deficit is noticed because the demands for inventiveness are so much greater at this time. At any rate, realization that the more conspicuously creative abilities appear to be concentrated in the divergent-thinking category, and also to some extent in the transformation category, we now ask whether we have been giving these skills appropriate exercise. It is probable that we need a better balance of training in the divergent-thinking area as compared with training in convergent thinking and in critical thinking or evaluation.

The structure of intellect as I have presented it to you may or may not stand the test of time. Even if the general form persists, there are likely to be some modifications. Possibly some different kind of model will be invented. Be that as it may, the fact of a multiplicity of intellectual abilities seems well established.

There are many individuals who long for the good old days of simplicity, when we got along with one unanalyzed intelligence. Simplicity certainly has its appeal. But human nature is exceedingly complex, and we may as well face that fact. The rapidly moving events of the world in which we live have forced upon us the need for knowing human intelligence thoroughly. Humanity's peaceful

pursuit of happiness depends upon our control of nature and of our own behavior; and this, in turn, depends upon understanding ourselves, including our intellectual resources.

References

- CHRISTAL, R. E., Factor analytic study of visual memory. *Psychol. Monogr.*, 1958, 72, No. 13 (Whole No. 466).
- GUILFORD, J. P., The structure of intellect. *Psychol. Bull.*, 1956, 53, 267-293.
- GUILFORD, J. P., *Personality*. New York: McGraw-Hill, 1959.

NANCY BAYLEY

Laboratory of Psychology

National Institute of Mental Health

On the Growth of Intelligence*

Recent studies have served to correct the former belief that man's intelligence did not increase with age. The type of study used is called *longitudinal*. It is illustrated here in the work of Nancy Bayley, who has studied the same group of individuals over a period of twenty-five years. Such studies, she points out, have produced fairly convincing evidence that intelligence continues to develop beyond late adolescence and that, although the rate of growth is slower, this development may continue well into our late adult years. She points out how we were misled by results of *cross-sectional* studies. If the student understands the differences between longitudinal and cross-sectional studies, he can see how the method used in collecting data can itself radically affect the results and the interpretation of results.

A central question dealt with in Bayley's work concerns the

* Reprinted and abridged with the permission of the author and the American Psychological Association from the article of the same title, *American Psychologist*, 10 (1955), 805-818.

"constancy" of intelligence. Do we remain at about the same level of intelligence throughout our lives? The concept of intellectual development would seem wholly incompatible with such "constancy"; Bayley surely provides evidence to throw doubt on the idea. However, in order to make reliable predictions of intelligence, there must be some constancy or regularity. This should direct the student to the study of individual growth curves (see pp. 469-472). With regard to each other's growth curves, children maintain relatively stable positions. For example, the distinctly superior and inferior children seem to keep their respective positions although development is present in both cases. Bayley also points out that rates of growth, measured over long intervals of time, remain fairly constant. However, measured over shorter intervals, growth curves can spurt ahead, level off, and spurt forward again. Other curves indicate only slow but steady gains.

For reviewing this article, the following questions are provided: (1) What characteristics of longitudinal studies and cross-sectional studies result in different data and conclusions? (2) In interpreting her data, why did the investigator find a "group factor" theory of intelligence more useful than a general factor theory? How would you interpret her data using Burt's concept of intelligence? (3) How can information on the development of intelligence be used by the classroom teacher?

The Selection of Infant Tests

When the Berkeley Growth Study started in 1928, we searched the literature for descriptions of infant behavior that would be suitable for evaluating intellectual development during the first year. The list tentatively compiled for our mental tests was heavily loaded with items from Gesell's norms, in their first formulations as published in 1925. Many of these items were closely similar to those listed in other sources, but Gesell had assembled an excellent set of materials on which to test these behaviors. He was also one of the few who had actually tested a fair sample of infants, thus furnish-

ing good preliminary norms. We selected from both the test materials and test items as first described by Gesell, adding items from other sources. In many instances we found it necessary to work out our own standard procedures and criteria of success or failure.

These tentative schedules we applied to the 61 babies of the Berkeley Growth Study, each infant being brought in at monthly intervals, starting at approximately one month of age. Ratings and descriptions were made on each child's responses during the testing situation. The items finally included in the California First Year Mental scale were selected after analysis of their adequacy according to the usual criteria. These criteria include: their occurrence in all or most of the infants; the increasing percentage of success on them with increasing age, for appropriate developmental stages; their internal consistency and correlation with the total behavioral criterion; and their apparent relevance as intellectual, or adaptive, functions.

Prediction from Scores in Infancy

At the outset we had accepted the findings based on school age children, and assumed that IQ's were constant at all ages. Consequently we were amazed at the precocity of some of the babies whose mothers seemed not very bright, and embarrassed at the poor records of other babies who, by the laws of inheritance, should have done better. But we soon found that our embarrassments and amazements were alleviated with time: a slow baby would forge ahead and redeem his inheritance, a precocious infant often seemed to rest on his laurels while the others caught up with him. We were not too surprised, therefore, when the statistical treatment of the test scores revealed that there was no relation between relative performance in the first few months of life and scores earned at the end of the first year.

When the report on the mental scores of the Berkeley Growth Study children during their first three years was published, it was met by many with scepticism. However, in spite of their failure to conform with established theory, these Berkeley children continued to develop in their own individual ways. What is more, we have corroborative evidence from the records of the children in the

Guidance Study, as reported by Honzik, and more recently by Honzik, Macfarlane, and Allen. Furthermore, these irregularities in mental growth were found to occur in other than Berkeley children. Wherever careful statistics have been applied to comparisons of repeated test scores on infants and very young children the correlations between tests separated by a year or two are low. It is now well established that we cannot predict later intelligence from the scores on tests made in infancy. Scores may be altered by such conditions as emotional climate, cultural milieu, and environmental deprivation, on the one hand, and by developmental changes in the nature and composition of the behaviors tested, on the other. These latter factors are the primary concern in this paper.

As the Berkeley Growth Study children grew older we continued to record their progress by successive tests at frequent intervals. We have from time to time reported the results of these tests, along with efforts to find relationships between mental growth and other factors. When the children were 8 years old a study of the individual growth records showed that only a fifth of the group had maintained any stability in their relative status over the eight-year span. Even these few had unstable Standard Scores during the first two years.

This lack of stability in infant test scores has resulted in various efforts to supplement and to correct the infant tests to make them more predictive. It has been suggested that the scales are not composed of the right kind of test items. However, efforts to devise other, more adequate scales, invariably run into the hard fact that infants exhibit a very limited range of behaviors that can be observed and recorded. The various scales of infant intelligence have a remarkable similarity of content. At first there is little to note beyond evidences of sensory functioning in reacting to appropriate stimuli. One can observe that the one-month-old looks momentarily at a dangling ring, or at a rattle or other small object. Or one can vary the source of the sharp sound that will make him start or blink. A little later the responses are evidenced in motor coordinations: the six-month-old may pick up a one-inch cube or a teaspoon placed in easy reach. There are some early evidences of adaptation to the presented stimuli, of memory from a past experience: the seven-month-old, for example, looks "aware" that a fallen toy is no longer there, and when a little older he may turn to look for it on the

floor. One can note the progression of vocalizations as they become more complex and then as they are used meaningfully. There is a developing ability to discriminate differences, to be aware of new situations, to recognize differences between members of the family and strangers, and so on.

The question is: Which, if any, among these is the forerunner of later intellectual functions? Which, if any, will predict the individual differences found in school age children?

One method of testing and selecting predictive items has been to use a later (or "terminal") measure of intelligence as the criterion. Scores earned by infants or very young children on individual test items have been correlated with their later IQ's. Those items showing the highest r 's with the criterion have in some instances been combined into scales. Theoretically, if other items of similar nature are then devised and added, such a scale can be expanded into an adequate predictive test.

We have tried to find predictive items from the First Year Scale on the Berkeley Growth Study children. Several years ago, using the six children at each extreme of intelligence as measured at the 14 to 16 year tests, we went through the First Year Scale item by item, noting the age at which each of these 12 children first passed each item. We were able to select 31 items in which the six high-scoring teenagers had, as infants, been two months or more advanced over the six low-scorers. These items were an odd assortment, and there was no evident reason for their superiority over other items. Most of the items occur in the second half year, where there is a fair amount of range in scores. In the first few months very few items had a range of more than two months in age at first passing.

Recently we computed scores for the total Berkeley Growth Study sample on this 31-item scale for three ages: months 6, 9, and 12. The r 's of these new point scores with the mean of the intelligence sigma scores at ages 16, 17 and 18 years (for 45 cases) are .09 at six months, .32 at 9 months, and .30 at 12 months. We were unable to get significant correlations even though our sample was composed in large part of the cases on whom the items were selected, including all of the extreme cases that would determine a relationship.

So far, none of these efforts has been successful in devising an intelligence scale applicable to children under two years that will predict their later performance. The moderate successes of Maurer

and of Hastings have been on items at the two-year level of difficulty or older. Even here the r 's are not high enough for accurate prediction on individual children. As far as I know, no one has used these items to set up and standardize an expanded scale. There does seem to be some coherence in the types of function tested by the predictive items. It is interesting to note, too, that those items which are good predictors are often not the items that best characterize a child's current stage of development. It has even been suggested that a scale should combine both types of items and then be scored in two ways—one score for evaluating present status and one for predicting future development.

These findings give little hope of ever being able to measure a stable and predictable intellectual factor in the very young. I am inclined to think that the major reason for this failure rests in the nature of intelligence itself. I see no reason why we should continue to think of intelligence as an integrated (or simple) entity of capacity which grows throughout childhood by steady accretions.

The Changing Organization of Intellectual Processes

Intelligence appears to me, rather, to be a dynamic succession of developing functions, with the more advanced and complex functions in the hierarchy depending on the prior maturing of earlier simpler ones (given, of course, normal conditions of care). The neonate who is precocious in the developing of the simpler abilities, such as auditory acuity or pupillary reflexes, has an advantage in the slightly more complex behaviors, such as (say) turning toward a sound, or fixating an object held before his eyes. But these more complex acts also involve other functions, such as neuro-muscular coordinations, in which he may not be precocious. The bright one-month-old may be sufficiently slow in developing these later more complex functions so as to lose some or all of his earlier advantage. This is the kind of thing that does seem to happen. Scores on tests given a month apart are highly correlated, but the longer the time interval between these baby tests the lower the interest correlation.

If intelligence is a complex of separately timed, developing functions, then to understand its nature we must try to analyze it into its component parts. One approach to this process has been by factor analysis. Of the two main theories resulting from factor analysis, our

data would seem to fit better into some variation of a multiple-factor than a two-factor theory. Or perhaps they fit better a theory that is intermediate, somewhere between the two.

The program of the Berkeley Growth Study has not been carried on in such a way as to make factor analysis on this material practicable. For one thing, the number of cases is too small for the usual factorial procedures. Also, for such a purpose one might have chosen a different or a more extensive series of tests. (As it is, the children have tolerated an amazingly large amount of testing and measuring!)

Nevertheless, some of our findings should point the way to new areas where factorial or other kinds of analysis would be fruitful. I should like to know, for example, where to look for g in the infant scales. One might expect g to be that factor on which prediction could be based. If g is not present at first, then when and how does it appear? Or does g itself change as it grows more complex? How do factor loadings distribute themselves in infant scales? Does a heavily-loaded first factor show a characteristic developmental process of change?

Richards and Nelson, using the Gesell items, at 6, 12, and 18 months, obtained two factors which they called "alertness" and "motor ability." They found age changes in communality of the tests that were in part due to restrictions in the type of items included in the scales at the older ages. This very fact reflects the relatively undifferentiated nature of behavior in the very young. It may be a mistake to try to call any infant behavior before 6 months more characteristically "mental" than, for example, motor. In spite of progressive selection of behaviors observed in intelligence tests, the evidence of a motor factor persists in the early ages of the Stanford-Binet, according to McNemar's factor analysis. These studies only scratch the surface of what needs to be done to gain real understanding of the nature of early mental processes.

If the word "intelligence" is best used as a broad general term that we apply to a great variety of mental functions, then we will want to investigate the nature of these functions, their interrelationships and the changes that take place in mental organization with growth. We should expect a given "factor" of intelligence to be more important at one stage of development than at another. As Garret has shown, in a summary of factor analyses, there is evidence of increasing independence of mental factors as children grow

older. Does this trend continue indefinitely? Or do some of these factors become functionally reintegrated as they mature? The studies of Thurstone and others can be most valuable in yielding information on this point. Let us hope they will be continued, over the entire life span, with careful attention to the problem of selecting items to test all relevant mental functions at all ages.

The very fact that the scores of mental growth in individual children tend to exhibit gradual shifts in relative status supports the theory that a changing organization of factors is in process. Something akin to g , or a high first-factor loading, must appear soon after the second or third year. The correlations of tests at these ages become positive with the later test scores. After 5 or 6 years children can be reliably classified into broad categories of normal, defective, and bright.

Problems Encountered in Constructing Curves of Growth in Intelligence

The use of intelligence quotients, or standard scores, in studying growth changes in children is helpful in showing a child's shifts in status relative to the norms. But a child's progress, in relation to his own past, is better represented if we can use scores that measure in-

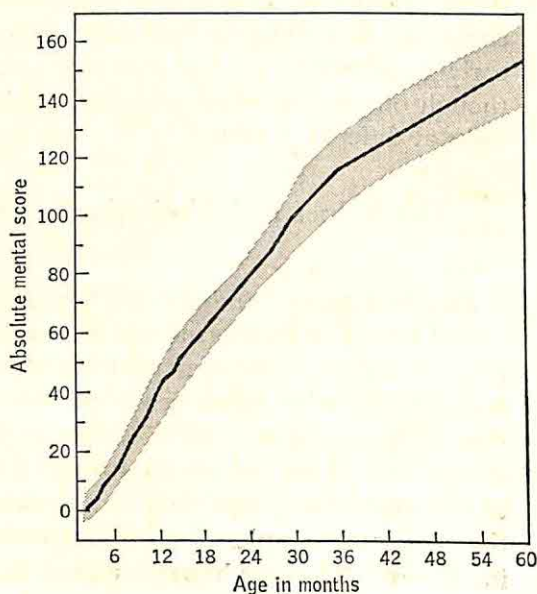


Fig. 1. Curve of Intelligence, One Month to 60 Months, Berkeley Growth Study, According to Thurstone's Method of Absolute Scaling.

crements or *amounts* of intelligence. Here we run into the problem of comparable units. Lacking absolute units for measuring intelligence, we must settle for some measure of greater or lesser difficulty, or degree of complexity of intellectual functioning. The first, and perhaps still most generally used unit of intelligence is mental age. Such a unit tends to force the same value on a mental age increment of (say) a month, whether it occurs at 6 months of age, at 6 years or 16 years. Thurston, Thorndike, and others have tried by various devices to set up units that approximate equality of difficulty at all levels of complexity. This is done usually by comparing the overlapping distributions of scores earned by children of successively older ages. Such units would vary with the test and with the normative sample. In any event, they remain only approximations. When we accept and label them as such, however, they become useful in comparing age changes in ability.

Thurstone applied his method to the Berkeley Growth Study scores on the California First Year and Preschool Scales for the first two years. We later extended the scaling through five years, and obtained the curve shown in Figure 1. This curve is positively accelerated for a few months, then settles into a consistent rapid growth for almost a year, after which there is a gradual slowing down in the rate, though growth continues to be fairly rapid. The curve makes sense in the light of ordinary observations of children's early development. It seems to be a useful approximation, even though one cannot claim absolute equivalence of difficulty of the units at different levels.

The Values and Limitations of Standard Scores and Increment Scores

Standard Scores: Individual Curves. Let us consider some individual curves of intelligence scores earned by these subjects. In the past we have usually presented individual records in the form of Standard Scores or Sigma Scores for this group. Such scores are very useful for observing a child's changes in performance relative to others his age. We can see his ups and downs, and try to relate them to variable factors, environmental or other, that might have caused the changes. Examples of an individual's relative curves are shown in Figure 2. Here we have presented, for the same child from birth

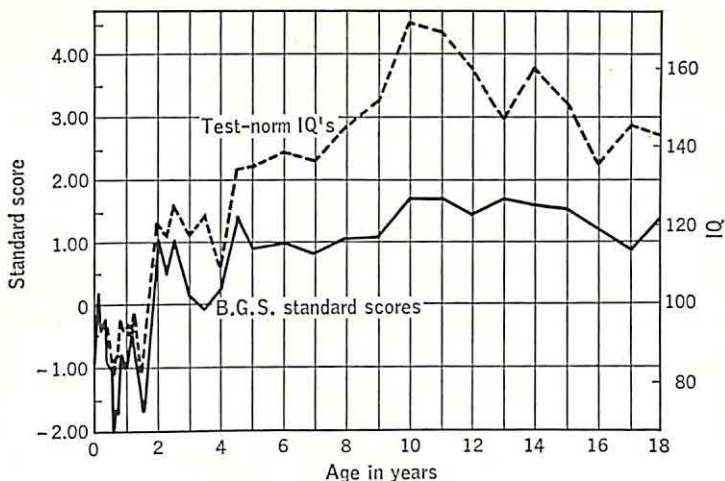


Fig. 2. Standard Scores and IQ Curves for Case 5M. From Bayley, N. Consistency and variability in the growth of intelligence from birth to eighteen years. *J. genet. Psychol.*, 1949, **10**, 188.

to 18 years, his IQ's and his Standard Scores. We have found that the Standard Scores gave a truer picture of a child's relative status at successive ages, because there were age changes in variability of the IQ's. That is, the variability of the M.A. did not, as had been assumed, increase with age in such a way as to maintain a constant *SD* of 15 or 16 points in the MA/CA ratio. We see in Figure 3 that IQ *SD*'s were greatest at one month and at around 10 to 13 years in this sample; they were least at one year, with a secondary restriction of variability around 6 years. In Figure 4 we have a child who was precocious early, but developed slowly.

Intercorrelations. In our comparisons with such things as emotional and environmental factors that could affect test scores, we have found the Standard Scores to be of value. For example, we have correlated the children's Standard Scores on intelligence at successive ages with the amount of schooling achieved by their parents. The age-changes in correlation (as expressed in *Z* scores) for this comparison are shown in Figure 5. The infant's scores at first are independent of parental status or negatively correlated, but after 18 months the *r*'s become positive, and by 5 years are about .55. In-

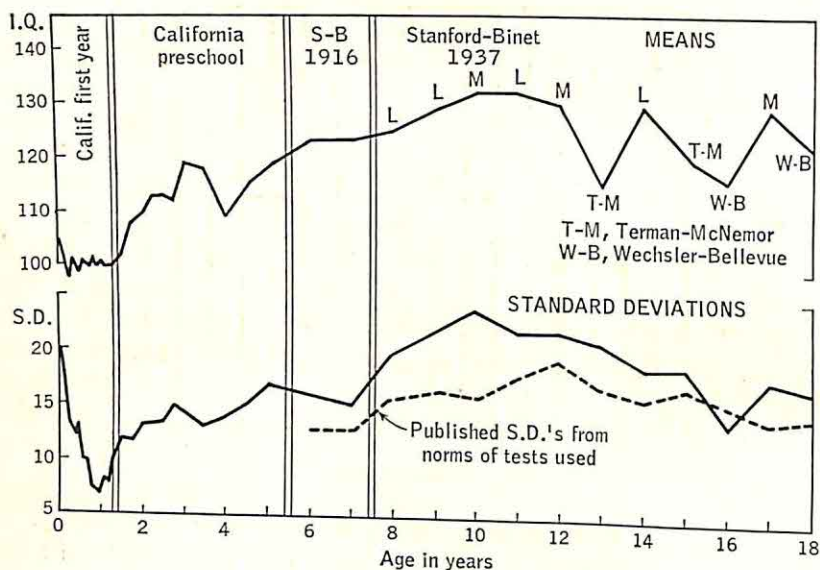


Fig. 3. Curves of the Means and SD's of IQ's, One Month to 18 Years, Berkeley Growth Study. From Bayley, N. Consistency and variability in the growth of intelligence from birth to eighteen years. *J. genet. Psychol.*, 1949, **10**, 171.

dividual curves in Figure 6 illustrate differences in the ages at which children's scores approach the level of their parents' educational status as expressed in standard scores.

Standard scores have been used to correlate mental ability with emotional factors. For example the r 's between children's standard scores and the amount of time they spent crying during the period of observation and measurement were at the zero level during the first year. Then, too, the repeated standard scores obtained for one child on intelligence, can be correlated with repeated scores on other variables, using the repeat observations on a single child as a population. For example, I obtained an "Optimal" score for each testing by combining 8 ratings that were indicative of the babies' responsiveness, or attitudes that might affect their performance on the tests. The r 's between these Optimal scores and intelligence at any one age were close to .30. Twenty of the children had Optimal

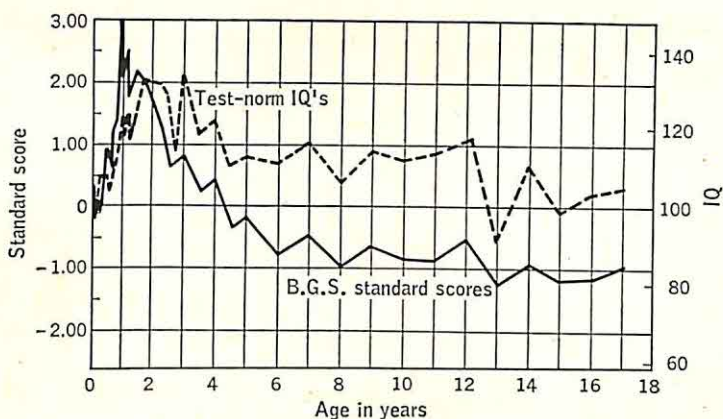


Fig. 4. Standard Score and IQ Curves of Case 14F. From Bayley, N. Consistency and variability of intelligence from birth to eighteen years. *J. genet. Psychol.*, 1949, **10**, 188.

scores available for from 12 to 15 test ages each, for the age-span between 6 months and 3 years. Using the rank difference method of correlation, rho's were computed for each child between his mental standard scores and his corresponding Optimal scores. These rho's ranged from plus .77 to minus .33. For similarly constructed "Attitude" scores, based on ratings made between 2 and 7 years of age, the individual children's rho's ranged from plus .76 to minus .46.

The wide range of correlations obtained corroborates the impression that observable emotional factors and attitudes (seen also in age curves of the different variables), rated at the time of the test, are to some extent related to the test scores, and evidently serve to help or to hinder the child's intellectual functioning. But other factors are also operative in determining a child's shifts in scores. These other factors may, in some cases, be so strong as to override the effects of emotional attitudes, resulting in negative correlations between mental performance and the child's observed responsiveness to the testing situation.

It becomes evident that the intellectual growth of any given child is a resultant of varied and complex factors. These will include his inherent capacities for growth, both in amount and in rate of

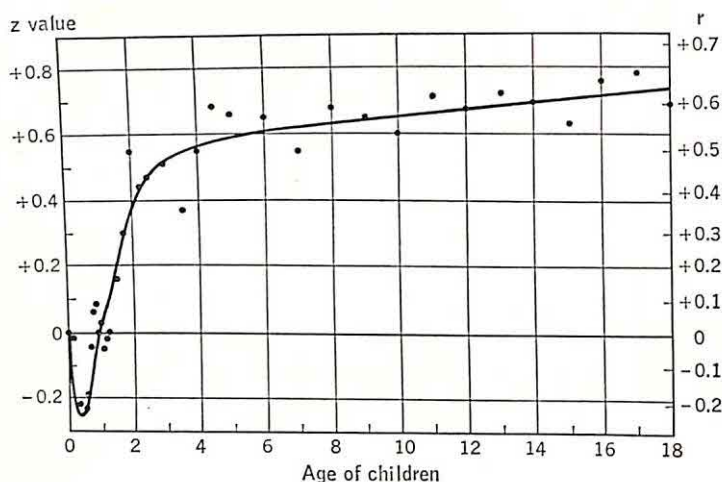


Fig. 5. Correlation between Children's Intelligence Scores and Parents' Education. From Bayley, N. Development and maturation. In H. Helson (Ed.), *Theoretical foundations of psychology*. Princeton, N.J.: Van Nostrand, 1951, p. 7.

progress. They will include the emotional climate in which he grows: whether he is encouraged or discouraged, whether his drive (or ego-involvement) is strong in intellectual thought processes, or is directed toward other aspects of his life-field. And they will include the material environment in which he grows: the opportunities for experience and for learning, and the extent to which these opportunities are continuously geared to his capacity to respond and to make use of them. Evidently all of these things are influential, in varying amounts for different individuals and for different stages in their growth. Many of these factors can be studied by observing concomitant variations in Standard Scores.

Individual Differences in Growth Rates

But Standard Scores, and other measures of relative status, have limited usefulness in the study of individual differences in rates of growth. Relative scores tend to make us forget that intellectual growth is a dynamic ongoing process, in which both averages and

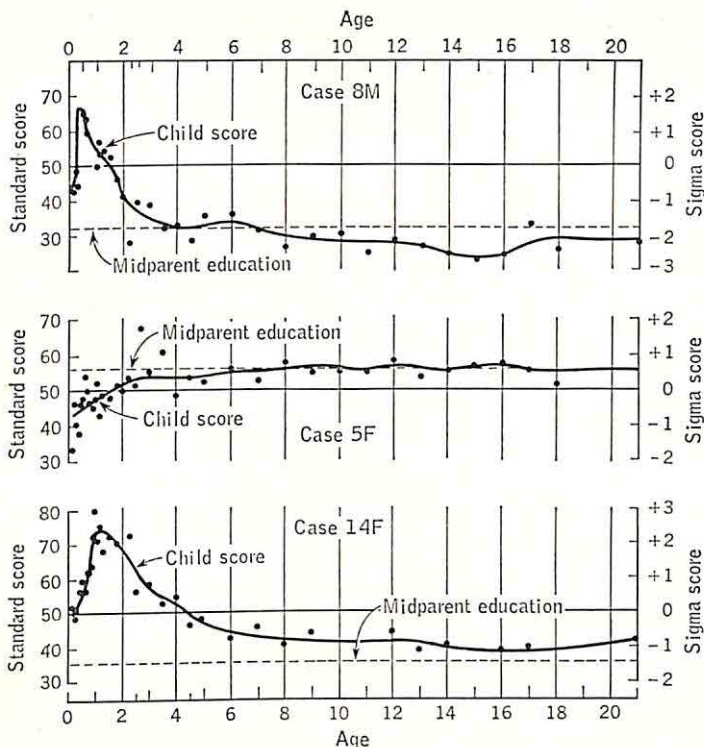


Fig. 6. Individual Curves of Intelligence Scores, Showing Differences in Ages at Approaching Parents' Educational Level. From Bayley, N. Development and maturation. In H. Helson (Ed.) *Theoretical foundations of psychology*. Princeton, N.J.: Van Nostrand, 1951, p. 17.

standard deviations in scores are related to the age of the subjects. It is worthwhile, therefore, to try to present individual curves of growth in units that will emphasize a child's change in relation to himself. Growth curves will enable us to observe a child's periods of fast and slow progress, his spurts and plateaus, and even regressions, in relation to his own past and future.

Such a growth curve has been shown in Figure 1, based on absolute scale units, for the first five years. In Figure 7, I have added two individual curves, superimposed on the curves of the mean and *SD*'s.

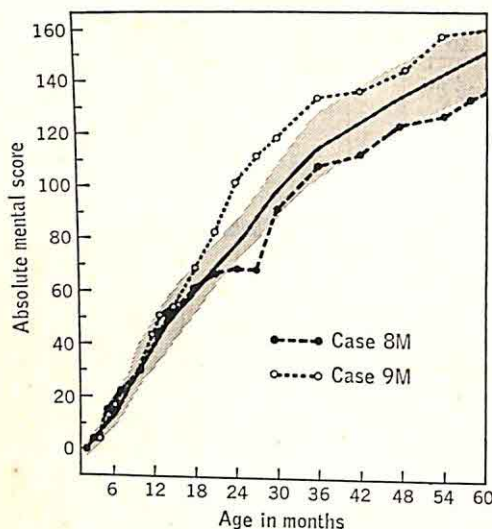


Fig. 7. Individual Curves of Growth of Intelligence in Absolute Scale Units, Showing Contrasting Patterns. From Peterson, H. A., Marzolf, S. S., and Bayley, N., *Educational psychology*. New York: Macmillan, 1948, p. 31.

Both of these boys tend to score above average during their first 18 months. Then case 9M becomes outstandingly superior for a year or two, while 8M suddenly lags behind. Study of the complete sample of individual curves reveals a great variety. There may be plateaus, periods of no growth, and occasionally actual decrements. There may be rapid forging ahead. Each child appears to develop at a rate that is unique for him.

By using the 16D scale we are now able to construct individual curves that extend for the entire period of the study. Figure 8 gives 16D curves for 5 boys. They cover the age span from one month to twenty-five years. Compared to the later years, the first five years seem very reduced in scope and the curves appear very homogeneous. If they are expanded to the same scale as the Thurstone curves, we find that for any given child both curves show the same periods of acceleration and retardation. The slopes of the Thurstone and the 16D curves are somewhat different, but the patterns of accelerations and retardations are generally similar in nature.

Although each child has his own individual pattern of progress, the patterns are not completely random. After the period of infancy there is a strong underlying consistency or constancy. Some children forge ahead and maintain relatively advanced positions

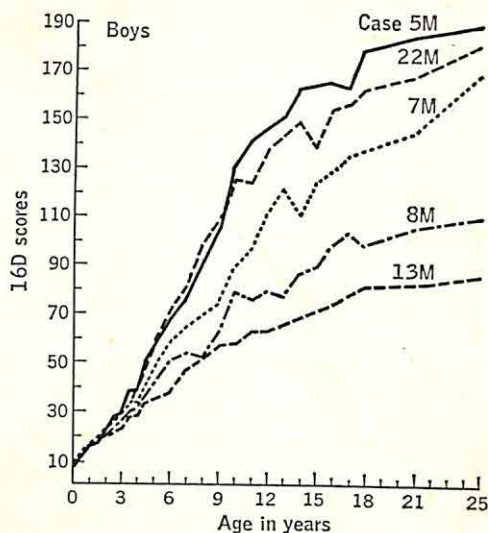


Fig. 8. Individual Curves of Intelligence (16D Units) for 5 Boys, One Month to 25 Years. Berkeley Growth Study.

after 5 or 6 years of age. Others grow slowly and lag behind. There is some shifting of position, but the changes are gradual over rather long intervals of time. Within such intervals we can expect to obtain fairly constant Standard Scores (IQ's).

It is notable that these five boys have all been tested at 25 years, and all five have continued to improve in their Wechsler-Bellevue scores. The continued growth occurs at all levels of ability. Case 13M, the slowest boy, has had increments in his Wechsler-Bellevue IQ's from 63 at 16 years to 78 at 25 years. This boy spent much of his childhood (ages 10 to 23 years) in an institution for the mentally retarded. When tested at 21 years he had never learned to read more than a few words. Now at 25 he reads, slowly to be sure, but he read aloud without error the Wechsler-Bellevue arithmetic problems.

Similar 16D curves are shown in Figure 9 for five girls, four of whom have been tested at 25 years. All but one of the four gained in scores at the 25-year test.

Some of the dips in the individual curves are due to changes in the tests. For example, those who have trouble in reading make relatively low scores on the Terman-McNemar Group test. But often the irregularities cannot be attributed to changes in the tests used at different ages.

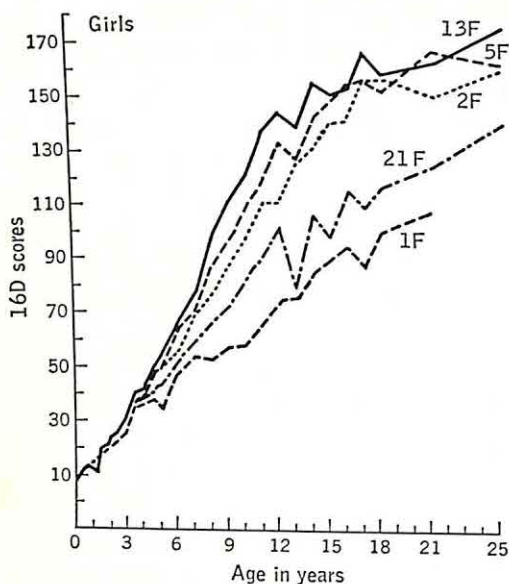


Fig. 9. Individual Curves of Intelligence (16D Units) for 5 Girls, One Month to 25 Years. Berkeley Growth Study.

Slight irregularities may reflect temporary conditions of motivation, health, or emotional factors. The more constant shifts require other explanations. Though they may result from prolonged emotional or environmental influences, they may also express inherent tendencies to develop at given rates. I suspect that each child is a law unto himself: in some instances certain factors are more important, while in others different factors play the determining role.

Temporal Changes in Adult Intelligence

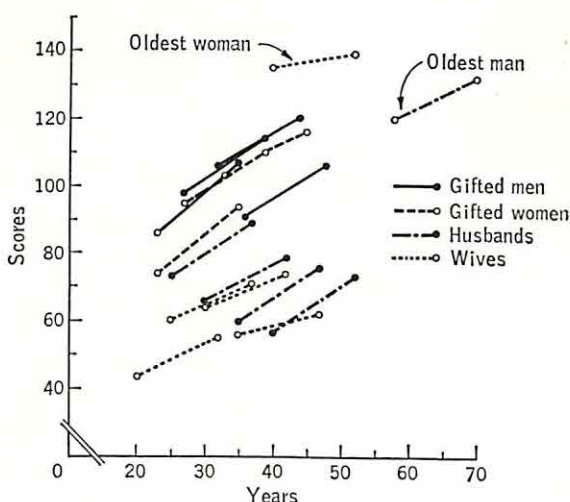
The few 25-year scores so far available indicate that the intellectual processes measured by these tests have not yet reached a ceiling. Fourteen out of fifteen subjects tested show continued increments. If these are typical cases, what, then, may we venture to predict for the years ahead? The alternative explanation of practice effects from repeating the same test might be offered. But the intervals between repeats on the Wechsler-Bellevue are 2, 3, and 4 years. These are rather long times to remember much about the specific items. Nevertheless, there is probably some residual memory for, or vague

familiarity with, the task and the type of solution found at the previous testings. At present we must assume that these factors account for part of the increment.

On the other hand, we have some recent evidence that some intellectual functions do continue to improve with age in adults. When the same individuals are retested after long intervals on the same test or on an alternate form of a test, the scores on the retests are significantly higher. These retests were carried out on superior adults, and their patterns of mental change may be different from those of less able persons.

In a recent study of the adult intelligence of the subjects of the Terman Study of Gifted Children, Bayley and Oden found that scores on the difficult Concept Mastery test increased on a second testing. For a population of over a thousand, composed of Gifted Study subjects and their spouses, comparisons were made between two tests that had been taken about 12 years apart. The increase in scores on the retest averaged about half a standard deviation. The subjects ranged in age from about 20 to about 50 years. When they were grouped into 5-year age intervals, the test-retest scores of all age groups increased, as is shown in Figure 10.

Fig. 10. Mean Scores on the Concept Mastery Test, Subjects Grouped According to Age at First Testing of Terman's Gifted Study Subjects. First and Second Tests for Each Group Are Connected. From Bayley, N., and Oden, M. H. The maintenance of intellectual ability in gifted adults. *J. gerontol.*, 1955, 10.



Similar results have been reported by Owens who repeated the Army Alpha test at 50 years on 127 men who had first taken the test as 19-year-old freshmen at Iowa State College. Their scores improved by .55 *SD*'s over the 31-year interval. One can hardly claim practice effects after a lapse of 31 years. Even the 12-year interval of the Terman study is rather long for any such claim: also the Gifted Study subjects were retested on an alternate form, thus ruling out specific memories of items. Furthermore, there were control groups consisting of those who were tested only once, at either the 1940 or the 1951 testing. The differences in mean scores of these groups at the two testings are the same as for the twice-tested groups.

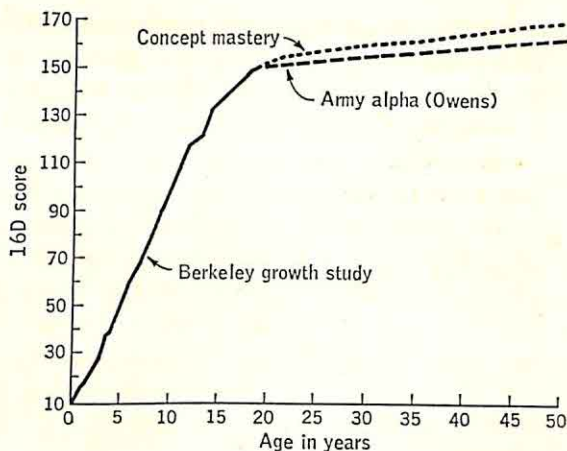
A Suggested Fifty-Year Curve of Intelligence

I have experimented with using the data from these two studies of adults to extend the 16*D* growth curve to 50 years. The subjects of the Berkeley Growth Study are, on the average, a somewhat superior group. Their 16-year Wechsler-Bellevue mean IQ is 117, and their 17-year Stanford Binet mean IQ is 129. A small group of 25-year-olds who have taken the Concept Mastery earned scores close to the average for the spouses of the Terman subjects at that age. We may assume, then, that this sample is rather similar to the Iowa State Freshmen and to the spouses of the Gifted Study subjects, in its general level of test performance. It has, therefore, seemed reasonable to join the data from the Berkeley Growth Study directly to the scores of either of the other studies, in extending the curve, as in Figure 11.

This joining of the curves has been done for the Iowa study simply by placing the 19-year initial point at 19 years on the 16 *D* curve and the 50-year point at the equivalent on the 16*D* scale of an increase of .55 standard deviations.

For the Gifted Study spouses the process was a little more complicated, but it has yielded a series of intermediate points, giving some indication of the probable shape of the curve. To obtain these points, I plotted a series of *SD* increment curves, placing the successively older ages at points on the curves of the younger groups in such a way as to take into account the growth already attained at any new starting age. From these series of overlapping curves, a

Fig. 11. A Proposed Age Curve of Intelligence, Birth to Fifty Years. Based on data from the Berkeley Growth Study, the Terman Gifted Study, and Owen's Iowa Study.



smoothed curve was drawn, and equivalent 16D scores were read off at 5-year intervals.

The resulting two-pronged curve for the 50-year span shows a more modest increment for the Alpha scores of the Iowa men. The Concept Mastery scores of Gifted Study spouses gain a full standard deviation, or about twice as much. Of course, since both of these curves are only approximations, neither may be more correct than the other. The differences are probably due, at least in part, to differences in the testing instrument. The Concept Mastery scale for one thing, has far more top than the Alpha, and allows for much greater expansion upward.

We have here evidence that tested intelligence, as measured by verbal concepts and abstractions, continues to grow when populations composed primarily of superior adults are retested. Intelligence may also continue to increase in the less bright. Certainly, the less favored members of the Berkeley Growth Study are still improving their scores at 25 years. What is more, in several other studies there is evidence that this phenomenon is not confined to individuals tested at the University of California Institute of Child Welfare. Freeman and Flory, for example, divided the children in their study on the basis of scores at 12, 13, and 14, into low and high scorers. At the later ages, 16 and 17 years, the low-scoring group was continuing to improve at a faster rate than the high-scoring group.

A recent study by Charles reports retest IQ's for 20 adults who had been diagnosed in childhood as feeble-minded. Their mean childhood Stanford-Binet, 1916, IQ was 58 and their mean adult Wechsler-Bellevue IQ was 81. Charles accounts for this difference in two ways: errors of diagnosis in childhood, and evidence from other studies that people who score low on the Binet test tend to make higher scores on the Wechsler. Similar explanations have been offered for similar findings in other studies. But a mean increase of 23 IQ points amounts to 1.5 *SD*'s of either of the tests used. This is a rather large shift to be attributed to test differences in restriction of scores, to regression phenomena, or to errors in the original test. All 20 individuals improved on the retest. It seems to me quite possible that these people did continue to improve in their mental ability.

There are many gaps in our knowledge of the nature of intelligence, and many questions remain unanswered concerning age changes in mental organization. In the curve presented in Figure 16, there remains an unanswered discrepancy between the adult portion and data for these ages presented by earlier investigators, who have found decrements in scores with increasing age after about 21 years. In the earlier studies some types of functions held up better than others. Owens found that those abilities that had held up best on the cross-sectional samples were the same ones that increased the most on his retests. The real difference between the conflicting findings seems to lie in the longitudinal as opposed to the cross-sectional method of obtaining scores for successive ages. In the former we have a constant sample whose life experiences, age for age, will have been similar in pervasive environmental conditions, such as wars, technological advances, and methods of education.

If, after taking adequate account of practice effects, the increases still remain, then the next question is to inquire into the nature of the tests, and the extent to which they measure intellectual abilities. Do such tests as the Army Alpha and the Wechsler-Bellevue, for example, measure intelligence in adults? Or do they tend to reflect continued experience in an increasingly enriched environment? Do the younger generations have more opportunity to develop their intellectual capacities than did their parents, or even their older brothers and sisters? Or are we just measuring the effects of in-

creasingly widespread informal education made possible by radio, television, and other modern means of communication?

If, regardless of the cause of the improved scores, they reflect actual degrees of competence outside of the testing situation, then these scores continue to have practical value. Another practical question is: What norms should be used in measuring deterioration resulting from brain injury, or from senescence? Perhaps it will be necessary to compare a present 50-year-old man's score with norms for, say, those who are 50 in 1954, rather than with 50-year norms for other decades.

What normal age changes should we expect in mental organization? The curve presented here is a composite. The forms of growth curves vary according to the functions measured. We should expect differences in the steepness of increment and decrement in growth curves of the different functions, and differences in the ages at highest efficiency. These differences have been found consistently in cross-sectional studies. The question raised here is whether more adequate studies, of the same individuals through time, will not show that the age of highest intellectual capacity is later than we thought, and that the decrements in abilities are, correspondingly, deferred.

This curve is offered as an alternative to previously published age-curves of intelligence. I should like to see it tested with further research that would refine, modify, and extend it into a more complete and accurate representation of intellectual changes over the entire life span.

References

- BAYLEY, NANCY, & ODEN, M. H., The maintenance of intellectual ability in gifted adults. *J. Gerontol.*, 1955, 10, 91-107.
- CHARLES, D. C., Ability and accomplishment of persons earlier judged mentally deficient. *Genet. Psychol. Monogr.*, 1953, 47, 3-71.
- DEARBORN, W. F., & ROTHNEY, J. W. M., *Predicting the child's development*. Cambridge: Sci-Art, 1941.
- FREEMAN, F. N., & FLORY, C. D., Growth in intellectual ability as measured by repeated tests. *Monogr. Soc. Res. Child Developm.*, 1937, 2, 116.
- GARRETT, H. E., A developmental theory of intelligence. *Amer. Psychologist* 1946, 1, 372-378.

- GESELL, A., *The mental growth of the pre-school child*. New York: Macmillan, 1928. Pp. 447.
- HASTINGS, HOMER, The predictive value of individual items in preschool intelligence tests. Unpublished doctor's dissertation, Univer. of California, 1952.
- HONZIK, M. P., The constancy of mental test performance during the pre-school period. *J. genet. Psychol.*, 1938, 52, 285-302.
- HONZIK, M. P., MACFARLANE, JEAN W., & ALLEN, L., The stability of mental test performance between two and eighteen years. *J. exper. Educ.*, 1948, 17, 309-324.
- MAURER, K. M., *Intellectual status at maturity as a criterion for selecting items in preschool tests*. Minneapolis: Univer. of Minnesota Press, 1946.
- MCNEMAR, Q., *The revision of the Stanford-Binet Scale: An analysis of the standardization data*. Boston: Houghton Mifflin, 1942.

Individual Differences: All Men Are Created Unequal

Introduction

For the teacher who is preparing for tomorrow's class it would be comforting to know that every cherubic first-grader is as able as every other cherubic first-grader. But most teachers, if they think about it at all, do not believe that students are very much alike. The teacher, however, frequently must behave in the classroom as if they were. And forgetting individual differences may arouse less anxiety than continual thought about ways of adjusting the instruction to meet the needs of each student. Whether or not we do anything about them, any one hour in the classroom shows us that wide differences do exist. We are forcibly reminded of the distinctive qualities of physical appearance. Differences in ability (even if we cannot quite define "ability") are sometimes just as striking. For example, we may admire Johnny's adroitness in grasping new ideas, the fluency of Mary's paraphrasing, and the keenness of Elmer's inductive thinking. Nor can we avoid observing, sometimes with furrowed brow, Tony's halting perusal of a simple sentence, Stella's difficulty in understanding basic addition, and Jerry's slow progress after tackling new materials.

Individual differences in ability do exist. Symonds has demonstrated this by using the following figures:*

* Percival M. Symonds, *What Education Has to Learn from Psychology*, 2nd Edition (Bureau of Publications; Teachers College, Columbia University, 1961), pp. 94-95.

Mental Age	Number of Children Age 6	Number of Children Age 9
13		1
12		4
11		11
10		21
9	1	25
8	7	21
7	24	11
6	36	4
5	24	1
4	7	
3	1	
	100	99

We shall assume that the 100 six-year-olds are in the first grade. What do these data show? If mental age is a figure that shows how intelligent these children are (at least as measured by an intelligence test), we see that about one third (36 percent) are as intelligent as we would expect; they have a mental age of six. However—much to the surprise of many parents—64 percent of the six-year-olds are either above or below the average mental age for their own group. Assuming that the work required only the mental ability of the average six-year-old, about one third (1% + 7% + 24%) might find the first grade too easy (we do not know that they do). Similarly, about one third might find it too difficult. The situation is about the same for the nine-year-olds, except that there is a greater spread of mental age here. One nine-year-old is an average thirteen-year-old, while another is only as intelligent as the average five-year-old—an eight year range in mental age for the group. The range for the six-year-olds is six years. About 8 percent of the six-year-olds could be doing third or fourth-grade work, while 37 percent of the fourth-graders are finding fourth-grade work difficult to do.

To summarize, these figures illustrate the following: (1) there can be a wide range of ability in any one grade or class; (2) the academic demands made in one grade may be suitable for only one third of the students; (3) that there is a sizable overlap in the ability of children of different ages and grades, especially the least able of the older with the most able of the younger; and (4) the range of individual differences becomes greater as we move up the grades.

Recently, both the general American public and the individual educator have been concerned especially with the third of the students who find the work at their grade level too easy. This is phrased as the problem of educating the "gifted." In the first article which follows, Vernon discusses this problem and points out the difficulties in identifying them, predicting their future academic success, and finding ways within present-day school practices and democratic ideology to promote their learning. Graphic evidence of these difficulties is furnished by Frankel in his study of high-ability students in the Bronx High School of Science.

Differences in ability is only one aspect of the problem, although probably the most important one. There are differences in physique, previous learning, family background, personality traits, sex, social class, race, and even culture. A school located in almost any large American city may have to carry on its educational activities while coping with this entire array of differences. The readings which follow deal with social-class and personality differences as they relate to ability and intellectual pursuits.

What can the schools and teachers do about the wide range and array of differences? This is the question which Vernon carefully considers for both American and English schools. As a general social and educational problem the question was raised by Burt in his discussion on the meaning and measurement of intelligence and its possible genetic determination. The social and cultural context within which education must find solutions to problems caused by variations in student ability was discussed in the Introduction to Chapter Seven (pp. 416-418). Vernon has more to say about this. Skinner states that teaching machines can have the effect of reducing individual differences and sensitively adjusting instruction for individual students (pp. 181-182). Differences in the effects of frustration were also referred to in a study by Waterhouse and Child (pp. 117-128). Taken together, the readings should point out the many aspects of the problem as it confronts the schools and at the same time indicate some broad criteria any solution must satisfy.

Relationship of Readings in Chapter 8

The key article in this chapter is the one by Vernon. The author combines a sound knowledge of the psychology of individual differ-

ences with a knowledge of the many practical problems faced in the schools. It is an excellent illustration of the translation of theory into practice. The report by Frankel, an attempt to discover what besides high ability makes for success in school, furnishes evidence to support Vernon's conclusions. The article by Schatzman and Strauss shows that linguistic differences, important in school achievement, are related to social-class differences.

PHILIP E. VERNON
University of London

Education and the Psychology of Individual Differences*

This article, written by an eminent English educational psychologist, is another outstanding example of how psychological theory can be translated into educational practice. However, there may be some disappointment that the author is not more prescriptive about what should be done in the schools to solve this problem. On the basis of theory, Vernon recommends only a general framework within which workable solutions may be found. The student can compare Vernon's translation of theory into practice with the opinions expressed by Kendler (p. 386) and Melton (pp. 22-27). The student can see how the use of theory to shape practice is often complicated by such extra-psychological factors as, in the words of Vernon, social prejudice and traditions, financial shortages, increasing birth-rates and other factors (see p. 484). This state of affairs often is as frustrating for the educational psychologist as for the teacher.

* Reprinted with the permission of the author and the publisher from the article of the same title, *Harvard Educational Review*, 28, No. 2 (1958), 91-104. References are omitted.

The article is clearly related to the previous chapter on intelligence because Vernon focuses his discussion on differences in ability. He uses Hebb's distinction between Intelligence A, or potential intelligence, and Intelligence B, or observable and test intelligence. This can be compared with the distinctions Burt has made, especially the one he attributes to Plato—the distinction between cybernetic and dynamic intelligence (p. 421). Vernon's related discussion on the constancy of the IQ recalls the discussion in the previous chapter by Bayley (pp. 472–477). The student should inspect the evidence which has led Bayley and Vernon to somewhat different conclusions. Vernon believes that there is moderate constancy and that there may be over-all consolidation of the child's particular level of ability by the time he enters the first grade. Vernon's explanation of this is in terms of the stimulation present or lacking in the home environment. The student should compare this explanation with previous readings in this book: (1) Ausubel's discussion on creating learning opportunities and interests (pp. 73–80); (2) Harlow's view of environmental stimulation (pp. 85–97); and (3) Bruner's discussion of competency motivation (pp. 263–265). Vernon's suggestion that the educational environment keep "just sufficiently ahead of each pupil's capacity to stretch his mind to the utmost" (p. 491) also recalls our discussions of motivation, particularly Bugelski's comments on the role of anxiety in learning and the report by Waterhouse and Child on frustration (pp. 84 and 117–128). This view, however, seems opposed to that of Skinner, who would try to keep learning well within the capacity of the student (pp. 177–179).

In discussing the question of developing special courses of study as a way of adjusting the curriculum to individual differences in both ability and interest, Vernon raises the question of intelligence as being a bundle of group factors—a theory endorsed by Guilford (pp. 435–456). Vernon makes a distinction between group factors and "special factors" (as defined by Spearman—see Burt, pp. 431–432) and shows how conventional classroom assignments and grouping can perhaps best adjust to such differences. It may be that the teaching machine and

programed learning may be most helpful in adjusting instruction to such special factors.

A central issue in this discussion is the question of discovering the academically able and predicting academic success. In connection with this the student should answer the following questions: (1) Why may admitting only students with high IQ's to college be a mistake? (2) Why must we often trust to luck in discovering the future genius? (3) Why must our predictions of future ability and academic success be short-term and flexible?

One of the most urgent and most controversial questions in education today is what kind of organization will encourage the fullest development of the varied mental capacities and inclinations of students. There are certain fundamental psychological principles, and some advances in recent years in the field of mental testing, which may help to guide our views on educational policies. Although the present system of allocating students to suitable courses of study in England is about as different as it could be from the system in America, and although I possess little direct knowledge of the American educational system, it would seem worth while to pool our experiences and the results of the tremendous amount of research that has been done, in an attempt, perhaps, to see our way a little more clearly.

One of the chief advocates of educational acceleration and scholarships for the gifted has been Professor S. L. Pressey of Ohio State University. Let us take as our starting point three of the main arguments that he puts forward. First, the education of the most talented is unnecessarily lengthy, and, since adults tend to reach the peak of achievement in most fields in their thirties or earlier, we are wasting much of their productive life. Second, acceleration does not involve hurrying or strain, but represents rather the natural pace for the more able. Wherever it has been tried it has worked successfully, and there is no need for it to lead to any social or emotional maladjustment. Nevertheless objections of this kind are continually being put forward, when the real reasons for doing nothing are the economic and financial consequences, or pure inertia or prejudice.

Third, a vast range of differences in abilities and interests exists among students who have had similar educational opportunities.

Now the problem is even wider than this. In general, pupils and students are so heterogeneous in their characteristics and potentialities that it might seem desirable to plan for each one an education uniquely suited to himself. However this is neither possible, nor desirable, for education implies not only individual development but also the training of different individuals to conform to society's patterns of intellectual and social norms. Thus there are positive advantages in educating diverse individuals in groups. Nevertheless, there must be some restriction of their heterogeneity, otherwise the educational process becomes inefficient and frustrating to the students as well as to the teacher. It would be absurd, for example, to try to train imbecile children and university students of atomic physics in the same group. To take a less extreme example: the tiny country school where one teacher copes with a very wide age and ability range performs many valuable functions; but there is no doubt, in England at least, that its educational efficiency tends to be below average.

Given, then, that there must be some reduction in heterogeneity, the psychologist would surely urge the following stipulations. Any grouping should be based on some characteristic which: first, is stable and enduring; second, can be accurately assessed; third, has a major influence on educational progress; and fourth, is acceptable to society. Difficulties arise because few characteristics, apart from age and physical handicap, meet these requirements. Age does largely, though of course far from completely, determine intellectual, emotional, and sensory-motor maturity; there are no difficulties of assessment, and it wins general acceptance. There is general agreement, also, that the deaf, the blind and partially sighted, and certain other physically handicapped groups should be segregated for special schooling under specially qualified teachers, though doubts arise in deciding what degree of defect requires such treatment. All other types of homogeneity seem to arouse intense controversy. Differentiation on the basis of sex, for example is rejected by most psychological opinion; yet it occurs in many European and a few American schools and colleges and is strongly supported by many parents and alumni. The explosive topic of race will not be considered, and religion will be omitted as having relatively little

relation to educational capabilities. Socio-economic class deserves some consideration.

Social class is fairly stable and assessable. We know that it is so closely associated with cultural level and with attitudes toward education that it has a marked effect on educational progress. In England, for example, where children are graded according to their suitability for advanced secondary education at eleven to twelve years of age by means of objective tests of intelligence and achievement, about three times as large a proportion of children of the white-collar classes pass the tests as do children of manual workers. Furthermore, social class gives one of the best indications of future achievement and adjustment in school. Havighurst and his colleagues have shown that, despite the greater social mobility in America, the social class of an American student's parents largely determines the stage he reaches in the educational ladder, and the kinds of courses he selects in high school. Thus there is some justification for the practice, common in Europe and not unknown in the United States, of having separate secondary schools for the middle and upper classes, entry to which is mainly by parental income. This system, however, is repugnant to the temper of the age in all democratic countries. Unfortunately, class is so pervasive that almost any form of selection or grouping is likely to be affected by it. Thus, selection by ability has become a source of acute political dissension in England between the more conservative who wish to retain it and the socialist or working class who wish to abolish it and substitute something more like the American common, or comprehensive, school.

Yet another, apparently sensible, criterion for the grouping of pupils—namely, a common linguistic background, usually breaks down because it too is associated with racial or with social class differences. One important reason why upper class parents tend to segregate their children, though they would seldom openly admit it, is that they do not want them to talk like children in the industrial slums. The psychologist may deplore these sociological and political influences on education, but he cannot afford to neglect them.

Let us turn then to ability grouping. In the early 1920's, when group intelligence tests became widespread, it was all very simple. We were taught that the tests measured innate intelligence, unaf-

fectured by social background, and that this was the main determiner of educability. Thus if pupils could be put into groups which were homogeneous with respect to intelligence, each group could progress at its own rate and the teacher's burden would be very much lightened. Moreover, the IQ was supposed to give a reliable indication of a child's ultimate educational powers; whether he was bright enough, or too dull, ever to manage work of college level, or to tackle advanced topics like mathematics. To a large extent this advice given by psychologists was put into practice in England. Age grouping gave way to ability grouping, though it was soon realised that too wide an age range of bright youngsters and old dullards in a single class is socially unhealthy. The usual practice, then, came to be the classification of children within an age grade in any large school into three or more streams or tracks on the basis of intelligence, previous achievement, or a combination of the two. Also, as mentioned earlier, the brighter children were segregated from the average and duller ones at eleven years for accelerated schooling.

Nowadays views on intelligence have been greatly modified. Indeed the term has been largely discarded by many American psychologists, as too liable to misinterpretation. Nevertheless, the distinction between intelligence and attainments still has limited value, despite their close overlapping. Intelligence can be defined as the more general thinking capacities: capacities for reasoning, for grasping relations, for comprehension, for new learning, and for concept development. That is, capacities which are not so much specifically taught as picked up by children in the course of their interaction with the home, school and wider environments. Whereas attainments refer rather to concepts and skills which depend more on direct instruction and on the child's interest and industriousness in the particular subjects studied. Such a distinction is one of degree rather than of kind, and it is entirely false to think of intelligence as causing, or making possible, attainments. It is at least as arguable that, through the acquisition of attainments at home and at school, the child is enabled to build up his intelligence.

Both depend to a considerable extent on innate potentialities or maturation—on what the Canadian psychologist Hebb calls Intelligence A—some characteristic of the central nervous system which enables certain children to develop mentally, to form percepts and

concepts, habits and ideas, to build up complex intellectual skills more readily than other children. Intelligence A is the capacity to acquire Intelligence B—that is, the intelligence which we actually observe in everyday life or at school, and which our tests sample fairly effectively. Intelligence B, as Jean Piaget also shows in his post-war books, is built up gradually; it does not depend solely on the child's genes, but also on the stimulation of the child by the world in which he is reared and which gets him to exercise his potentialities. How far specific attainments involve different genes from those underlying general intelligence seems to me a moot point; one could certainly make a good case for it in the field of music and possibly other talents. But, like most heredity-environment controversies, this seems an unprofitable argument, since we can never in fact observe or measure Intelligence A or other potentialities directly; they are purely hypothetical constructs.

We can no longer regard intelligence as setting a fixed rate of mental growth, or as having a definite termination. The early pioneers of mental testing seem to have regarded the IQ as a fairly constant characteristic of each individual because there was good agreement or correlation between two applications of the Stanford-Binet test a week or a month apart. But this tells us little about the fluctuations in ability to be expected during five or more years of schooling. On surveying the many investigations that have been made into this problem recently, a number of technical snags were found. Most of the published results were more or less distorted by such factors as abnormal standard deviation of the IQs, extreme similarity or dissimilarity of the tests employed, above average intelligence level of the tested group, frequency of retesting, etc. However, allowing for these it appeared safe to conclude that either over the six to ten or the eleven to eighteen year period, the correlation between two similar (though not identical) intelligence tests does not drop below a coefficient of 0.70. This implies that the typical individual would vary only 7 IQ points up or down.¹ Although the majority do not fluctuate very widely, as many as 17 per cent might alter considerably, that is, anywhere from 15 points (or 1 standard deviation) up to about 40 points.

¹ The statement assumes that such tests are standardized to the same mean of 100 and standard deviation of 15.

All-round educational attainments seem to be at least as stable. Certainly it is a mistake to think of the IQ as something fixed, and attainments as varying widely with the teaching received and other environmental causes. I would suggest that both show moderate constancy, partly because both are based on genetic potentiality—Hebb's Intelligence A—but also because their growth is essentially cumulative. By the time the child reaches school at five or six years, the interaction of potentiality and home stimulus have consolidated in him a certain level of ability which will closely determine his rate of progress for the next few years. Good or poor teaching, or other marked environmental irregularities and changes in personality adjustment, motivation, and interests, will result in considerable alterations only among a small minority. Again by eleven or twelve years the consolidated level gives fairly close predictions of the rate of progress over the next six years, as we have been able to show by our follow-up studies of the English secondary school examinations. The correlations of intelligence tests, of English and Arithmetic tests, and of teachers' estimates of ability with successful secondary school performance over two, and even up to five years average close to 0.80, when suitably corrected for homogeneity. By combining these three sources of information the coefficient can be raised to about 0.86. This, of course, refers to a whole age group. The figure may sound surprising to American educational psychologists, because they normally do their testing within pre-selected groups such as students seeking college admission. But if they likewise calculated the efficiency of their aptitude and achievement tests and school grades in separating, say, the top 20 per cent of high school students who are most suitable for college courses from the bottom 80 per cent, there is no doubt that they would reach at least as good predictions.

Yet even a correlation of, say, 0.85 allows of a considerable degree of error. Suppose we did send 100 unselected eighteen- to nineteen-year-olds to college and found which twenty were the most successful, in fact only fourteen of the twenty would have been correctly indicated by our tests and high school grades. In other words, using the best criteria available, about one-quarter to one-third of those we should select for college are likely to be unsuitable. They would

have been equalled or surpassed by some $7\frac{1}{2}$ per cent of those thought unsuitable, had the latter been given the opportunity. We can represent the situation as follows:

	Successful in College	Unsuccessful in College	Total
Selected on the basis of tests and previous grades	14	6	20
Rejected on the basis of tests and previous grades	6	74	80
Total	20	80	100

This is precisely the situation that obtains in English selection for advanced schooling in what we call grammar schools. A very high statistical level of accuracy is reached, and yet many pupils admitted to grammar school are unsuitable, and some who are rejected turn out to be so-called late developers who have to be transferred later on, or otherwise show outstanding ability. The psychologist, however, would naturally expect such fluctuations in abilities during adolescence and young adulthood; as students progress to more advanced courses, their interests and adjustment often change and they develop unsuspected capacities, or drop below their previous promise. Whatever kind of system of selection or acceleration was adopted in America, really accurate prediction would likewise be found impossible.

Frequently one hears the argument that far too large a proportion of our talented youth fails to get to college; there are many millions with high IQs who are being wasted. This argument is no more impressive than the statement: since college students are above average in height, we should ensure that all six-footers reach college. While IQ may be somewhat more relevant, its correlation with success in college for a representative sample would not exceed 0.75, and this means that 40 per cent of students picked merely by high IQ would turn out unsuitable. Even if we raised our standards and insisted, say, that all students with IQs of 125 and above, the best 5 per cent only, should be accelerated in high school, or sent to college, nearly one-quarter of our choices would let us down. At the opposite extreme, even if we went as low as an IQ of 90 and

included the top three-quarters of the population, we would still find that 2 per cent of those excluded might have made the grade. In order to catch all or nearly all those capable of benefiting from a more advanced or accelerated educational course, we should need to go so far down in the scale of ability that many students would be brought in who would not benefit or who might have been more appropriately allocated to some quite different course, say of vocational training.

No definite conclusions or recommendations will be arrived at in this paper; but the main educational implication so far would seem to be that any predictions we make about students, or any selection, should be short-term and flexible. Such procedures should take into account not merely abilities but also interests and values—whether, for example, the student and his family are keen for him to take college preparatory courses or to go to college. While all assessments are likely to have a considerable margin of error, they will certainly be less inaccurate when based on the way the student is shaping and progressing in work of a similar nature to that which he will undertake. Clearly, too, caution should be exercised regarding any statements which imply that a considerable proportion of the population is always going to be intellectually incapable of absorbing the higher reaches of education. There is still a lot of truth in this, but the trouble is in determining *which* members of the population fall into this category. The limitation is at least as much a matter of defective intellectual interests and defective attitudes to education in the environment from which the weaker students come as it is to these students' innate intellectual inferiority.

Another deduction may be drawn from the work of Hebb, Piaget, and others on intellectual growth: such work suggests the desirability of the educational environment keeping just sufficiently ahead of each pupil's capacity to stretch his mind to the utmost. Hebb found that dogs or rats brought up in the restricted environment of a cage were less able, as adults, to perform new learning and problem solving tasks than those brought up in a richer and freer environment. Similarly with children, inadequate educational stimulation may mean not only that they learn less than they should, but that they also become less able to acquire further thinking skills. Experimental evidence can be provided on this point.

Recently, some 800 boys in a large English city who had been tested and allocated to fourteen very diverse secondary schools at age eleven were retested at fourteen. After allowing for initial level and regression effects, there were differences between the different school groups amounting to 12 IQ points. The pupils in the average selective or grammar school had gained 7 points (nearly half a standard deviation) over those in the non-selective schools. All this difference cannot be attributed to the effectiveness of schooling as such. Grammar school boys mostly came from better homes where they received more encouragement and help with their work, while the non-selective school boys lived in culturally and materially poorer homes, where there was often active opposition to education. Nevertheless, a rank correlation of 0.85 was found between the city administrator's assessment of the degree of stimulation likely to be provided by the fourteen schools, and the actual order of average gains among the pupils in these schools. Clearly, much more transfer had occurred than Thorndike found in his classic experiment on different high school courses.

This finding recalls Lorge's work on the intelligence level of thirty-four-year-old adults who had varying amounts of secondary and tertiary education after they were tested at fourteen years; and also a recent study by Husen of several thousand Swedish boys at nine and nineteen years. Husen found that those who had had full education throughout the period gained the equivalent of some 12 IQ points over those who had left school at fourteen to fifteen and received no further education.

During World War II, under the auspices of the British Navy, the Raven Matrices scores (a non-verbal reasoning test) of 90,000 men recruited at various ages from seventeen to forty were compared and a general tendency to decline with age was noted. Clearly, however, those who had been in unskilled and laboring occupations, where they had made little use of their brains, declined earlier and more rapidly than those from skilled trades and clerical work, which had presumably done more to exercise them.

Now according to the older views, intelligence was said to grow steadily in childhood, then to slow down and reach a constant maximum at around fourteen to fifteen years of age. No one ever explained how, if this were so, high school and college students could continue to increase their scores on suitable tests in the late 'teens.

But these results are quite reconcilable if we remember that, around the time of the World War I, probably the majority of average and duller individuals were leaving school by the age of fifteen; their intellectual capacities not only ceased to develop but began to stagnate, whereas the brighter, who either stayed on at school or entered more stimulating jobs, continued to improve. Thus the combination of a rising group and a declining group produced the apparently constant average level in the adolescent and young adult population.

The objection may be raised, as it was raised to Lorge's findings, that intelligence tests depend greatly on vocabulary and reading comprehension, capacities which are naturally much affected by length and quality of schooling. The objection rests, of course, on the old confusion between intelligence as a hypothetical innate potentiality (Intelligence A) and intelligence as directly manifested (Intelligence B). Actually, further work by Lovell, indicates that it is not mere verbal skills which are most affected by the stimulating or depressing qualities of adolescent environment, but rather the flexibility aspect of intelligence, or the capacity to form and apply new concepts. A more serious weakness is that we know so little, as yet, regarding the kinds of educational process that have greatest transfer value, or do most to stimulate the growth of ability.

Now the implications of such findings for educational organisation would seem clear. They definitely support some system like the English one of grouping pupils by ability, and pushing forward the brighter ones more rapidly in selective schools or classes. And they suggest that there is considerable substance in the complaint made by many American parents that the American system of public education seriously retards the bright child who is willing and able to tackle more difficult courses. On the other hand, there is the danger that average or duller students, if they are less pushed, will tend to fall more and more behind. Thus, it becomes even more difficult for those whose capacity happens to improve later to get back into the stream that would be appropriate. In other words, homogeneous grouping by ability tends to stereotype those who were less able initially and freeze them at a lower level.

This is a very real problem in English elementary schools, where pupils may be classified by teachers as early as seven years old into those thought likely to pass or fail the eleven year examinations.

Such classification, being based mainly on early progress in reading and number-work, naturally gives a great advantage to children from middle-class families who are more likely to be helped and encouraged at home. Thus, a rather rigid system of selection may mean that the initially slow learners drop so far behind that they are, in effect, being condemned very early to an inferior schooling and a manual occupation for life. However, evidence indicates that such rigidity is the exception rather than the rule; perhaps owing to the influence of psychological teaching, most primary schools now do retain a fair degree of flexibility of transfer from one stream to another, and a sufficient degree of overlapping between the curricula of the several streams.

A similar danger arises at the bottom end of the ability range, where, in many educational systems, the very backward pupils may be classified as mentally defective and sent to special schools. It may be that their abilities too become stereotyped at this level, although they might have been capable of returning to ordinary school classes later. Miss Bernardine Schmidt's famous research at Chicago definitely supports this view, although her claims are by now pretty thoroughly discredited. In England, at least, most mental defectives receive a type of education far better suited to their lowly intellectual level and their weak emotional adjustment than they would get in ordinary schools. Yet they usually stay as backward as ever or even decline. No one who has observed such children vegetating hopelessly in an ordinary school, and then busy and interested at a special school could doubt the desirability of some type of homogeneous grouping. Nevertheless, I agree that here too irreversible segregation is to be avoided, and that injustices do occur.

One other consideration that I wish to raise in respect to ability grouping is that—as educationists well know—any scheme for grading of students has a “backwash” effect on education, by acting as an incentive to the teachers, the students and their parents. The eleven year selection examinations in England provide a terrible object-lesson; many, though by no means all, primary schools concentrate so exclusively on cramming their fifth and sixth grade pupils for objective tests that any other educational activity, however valuable it might be to children's general growth, tends to get crowded out. Much ill-feeling is engendered among snobbish parents; often they

coach children themselves from published manuals of intelligence and other tests; or they communicate their anxieties to their offspring and sometimes induce serious strain. Actually the effects on young children's mental health are apt to be greatly exaggerated by the sensational press. A careful survey of child guidance clinic cases has been made and no tendency for an increase in referrals was found around the time of the examination. In only one child in about 500 does the examination appear to contribute to maladjustment, and then only in children who were already prone to anxiety through earlier upbringing or constitutional weakness. Probably other competitive examinations, such as those at the end of a secondary course, are associated with mental breakdown at least as frequently. Nevertheless, the existence of unnecessary stress even in a small minority, and distortion of the educational process among the majority, are the most serious defects of the English system, and have to be weighed very carefully against possible advantages in intellectual acceleration.

Now although sound psychological reasons for expecting intellectual benefits through grouping can be given, it is very difficult to prove. Furthermore, such limited experimental comparisons as have been carried out seem to have yielded negative results. It is doubtful whether such studies have been adequately designed to answer the important questions. International comparisons, again, yield no sure evidence. It is commonly stated that the products of English grammar schools, when they enter college, are two years ahead of American students of the same age. Even if this generalization were true it proves nothing, since only some 3 to 4 per cent ever reach the university in England. Also, it may well be that superiority in academic achievement is counterbalanced by poorer development in social and other more subtle qualities. It would seem, and some day it may be shown empirically, that the English system of grouping and selection does produce improvement in achievement at the top end, whereas the common school system is better for the average or dull, and also obviates many of the difficulties that arise through fluctuations in, or stereotyping of, mental growth. If this were so, it would be a matter for society rather than psychologists to decide which outcome it prefers.

No doubt it will be suggested that, instead of grouping by general intelligence and previous all-round achievement, we should consider type of ability and interests along special lines. This might be called the multi-dimensional as opposed to the uni-dimensional approach. British and American psychologists have always differed on this, just as have their educational systems. The British, following Spearman, have tended to stress the importance of general ability: children who are above average in one school subject are likely to be above average in all others. Only to a limited extent, as Burt demonstrated in 1917, do we get more specialised types of ability showing in the primary school—for example, some children being generally poorer at number work than at English subjects, or better at all practical and manual activities. In contrast, Thurstone in 1938, from his analysis of tests given to college students, claimed that abilities are largely independent—attributable to a series of separate factors; verbal, number, reasoning, spatial, etc. When he extended his testing downwards to eighth grade and younger children, who were less highly selected, he in fact found much overlapping among these factors. So that, while his statistical approach was quite different, his results confirmed Burt's in showing that something like 50 per cent of the variance in children's abilities, at least in abilities that are relevant to educational progress, could be accounted for by a general factor. This means then that it is extraordinarily difficult to differentiate secondary school pupils by type of ability into even such broad categories as linguistic or academic on the one hand, and scientific or technical-mechanical on the other. For the most part, those who are best at one course would also be above average at the other, and vice versa.

If we confine our attention to a narrow ability range, say the top 20 per cent as in England, then we find it possible, with our present verbal and spatial tests, to differentiate the extremes. Of course there are no distinct types of children any more than there are tall and short types in height. We can pick out roughly one-sixth who are relatively superior on the linguistic to the technical side, and one-sixth who are more suited to the other type of course. There remain two-thirds who are likely to do equally well or badly at either. Roughly the same is true of American college students. The efficiency of prediction of overall level of achievement by means of tests

is much greater than the efficiency for classifying according to, say, engineering or arts, and this is only the broadest and most obvious dichotomy. If we tried to go further and split off, say, foreign linguists, social studies students, mathematicians, biologists, physicists, engineers, commercial students, art students and agriculturalists, we should certainly be far less successful still. Probably the linguists, mathematicians and physicists would be distinguishable from the artists and agriculturalists more readily by their generally higher intelligence and previous attainment than by such specialised tests as we have available at present.

Some psychologists would claim that abilities differentiate with age, so that suitability for different lines of study or type of curriculum would become more clear-cut near the end of secondary schooling or during college years. No satisfactory evidence for this claim can be found except in so far as groups of older students are usually more highly selected in respect to general ability. In the independent college, the scientist might make a hash of French, or the good historian fail in engineering, because in such a population practically all cases fall within the top 10 per cent as regards general ability. There would be much less differentiation in the state universities, where students range roughly over the top 30 to 40 per cent of the scale.

However, the outlook for guidance or selection into types of courses is not quite so dim if we take account of interests, since interests have the tremendous advantage, for classification purposes, of showing quite low, sometimes even negative, correlations with one another. The good scientist and the good business man, for example, may differ little in such abilities as we are able to test. They would however loathe one another's jobs, and would likely be thoroughly bored by any education or training designed to prepare them for the other's technical career. In England we find that we can differentiate nearly two-thirds of the more able pupils into the academically and the technically-minded at eleven as against the one-third mentioned earlier, by taking account of *interests* as well as *abilities*. One difficulty with interests, though, is that they cannot develop without experience. There must be many adults who fail to get much satisfaction from their vocational and avocational pursuits because their education has not provided sufficient opportuni-

ties for them to experience, say, painting, farming, or mechanical work. The American system of offering a wide choice of secondary school subjects, while often criticised by educators for its fragmentariness, is sounder in this respect than the English which gives less variety and indeed tends to stereotype the more able pupils either in an Arts or a Science field from about the age of fourteen, and to discourage less academic inclinations. The shortage of technologists and high-grade technicians in England is at least as serious as that in America. Clearly one of the reasons is the high social prestige of academic courses of study which hinders the provision of experience that might stimulate the development of more practical interests.

Another obvious difficulty arises from the variability of interests during adolescence. Among many students interests may not be sufficiently stable nor assessable to allow firm and satisfactory educational or vocational specialisation until the end of secondary schooling or even later. Nevertheless, one investigation at London University of adult technologists and Arts men did suggest that some differentiation of interests would have been apparent in the great majority as early as twelve years. This is a field where psychologically trained counselors can already help in making earlier diagnoses, and where current work on personality gives promise of improved tests.

How about the problem of discovering the brilliantly talented individual—the occasional future genius? Clearly, neither our tests of abilities nor of interests are going to be much use here, and we must continue to trust mainly to luck, to the perceptiveness of the school or college teacher who encourages him, and to the individual's own drive, which enables him to break through the barriers of the educational system. Admirable as Terman's follow-up studies of children with high IQs are, some harm has been done by labelling them studies of "genius."

One other point of psychological theory: although it has been shown that the group factors underlying particular types of curriculum are disappointingly small and difficult to diagnose, this does not dispose of what Spearman called specific factors. Individual pupils show great variations with respect to specific topics within any course, depending on their particular past experience and interests, and on their emotional responses to the teacher and other

members of the class. Thus, even when they are satisfactorily grouped for, say, oral reading, they will show considerable heterogeneity in silent reading or spelling, let alone in non-related subjects. For such reasons, many educational psychologists currently look more to individual assignments and small-group work within classes to cope with most of the problems of individual differences than to overall grouping into separate classes. With this goes the notion of enrichment of courses among the brighter students. The adequacy of such solutions is doubtful since they clearly imply increasing the degree of heterogeneity within the class and thus run counter to the thesis of this paper.

What conclusions can we come to then? No evaluation of the various mechanisms of acceleration that have been, or are being, tried out in America such as grade-skipping, advanced standing, and the shortening of courses through enriched programmes will be made here. But the evidence does seem to point to the desirability of some form of grouping, such as indeed already exists in many American high schools. Nevertheless, we have seen that there are many dangers in introducing anything that implies competitive selection or stereotyping of ability levels. It is, therefore, preferable to keep to grouping by age, and later by interest, as far as possible. Up until about nine years of age or the fourth grade, there would seem to be no good case for any ability grouping other than segregation of the lower-grade feeble-minded, the physically handicapped, and, perhaps, temporary remedial classes for the higher-grade defectives and the very backward. However, by the age of about thirteen or the eighth grade, the range and complexity of abilities *appears* to have become much wider. It is still doubtful how satisfactorily we can measure range of ability in any absolute sense even at this age. But everybody can't go on studying everything, and, in the interests of professional and vocational objectives, some specialisation should begin to be introduced. Hence, some degree of grouping by general ability, and to an increasing extent by interest, would seem legitimate. While it is true that many thinkers strongly deprecate early specialisation, surely its dangers should be balanced against the waste of productive years which results at present from the late postponement of any specialisation. Between the fourth and eighth grades one can only suggest that the curriculum be largely of an

exploratory or diagnostic character, designed not merely to provide essential skills, but to stimulate general mental development and to provide experience out of which interests can be built. Even when grouping is introduced, there should be sufficient overlapping between groups to make transfer up or down easy, and of course sufficient common activities in a school or college to discourage the formation of barriers. In other words, the process should be one of gradual approximation in accordance with the principles of educational guidance rather than one of selection or irreversible decision.

Inevitably this is a vague kind of framework and the psychologist must always remember that his prescriptions are liable to be upset by social prejudices and traditions, financial shortages, increasing birthrates, and innumerable other factors. The American type of school organization seems to meet the prescription set forth in this paper more nearly than does the British, though it may, of course, show weaknesses in other respects. Some means should be devised of giving greater and earlier recognition to individual differences in general educability and of avoiding, if possible, some of the mistakes that have arisen in England as a result of the strangle hold of tradition and the well-intentioned but short-sighted policies of educational planners.

References

- BLANDFORD, J. S., Standardized tests in the junior school with special reference to the effects of streaming on the constancy of results. *Brit. J. educ. Psychol.*, 1958, 28.
- BURT, C. L., *The distribution and relations of educational abilities*. London: King, 1917.
- CLARK, J. F., A psychological and sociological study of students in higher technological education. Unpublished Ph.D. thesis, Univer. of London, 1951.
- HEBB, D. O., *The organization of behavior*. New York: John Wiley, 1948.
- HUSEN, T., The influence of schooling upon IQ. *Theoria*, 1951, 17, 61-88.
- LORGE, I., Schooling makes a difference. *Teach. Coll. Rec.*, 1945, 46, 483-492.
- LOVELL, K., Intellectual deterioration in adolescents and young adults. *Brit. J. Psychol.*, 1955, 46, 199-210.

- PIAGET, J., *The psychology of intelligence*. London: Routledge & Kegan Pal, 1950.
- PRESSEY, S. L., Educational acceleration: appraisal and basic problems. *Ohio State Univer. Bureau Educ. Res. Monogr.*, 1949, No. 31.
- THORNDIKE, E. L., Mental discipline in high school studies. *J. educ. Psychol.*, 1924, 15, 1-22, 83-98.
- THURSTONE, L. L., *Primary mental abilities*. Chicago: Univer. of Chicago Press, Psychometr. Monogr., 1938, No. 1.
- THURSTONE, L. L., and THURSTONE, T. G., *Factorial studies of intelligence*. Chicago: Univer. of Chicago Press, Psychometr. Monogr., 1941, No. 2.
- VERNON, P. E., The variations of intelligence with occupation, age, and locality. *Brit. J. Psychol. (Statist. Sec.)*, 1947, 1, 52-63
- VERNON, P. E., Intelligence and intellectual stimulation during adolescence. *Indian J. Psychol.*, 1957, 2, 1-6.
- VERNON, P. E., (Ed.), *Secondary school selection*. London: Methuen, 1957.
- WARNER, W. L., HAVIGHURST, R. J., & LOEB, M. E., *Who shall be educated?* New York: Harper, 1944.

EDWARD FRANKEL
Bureau of Educational Research
Board of Education of the City of New York

A Comparative Study of Achieving and Underachieving High School Boys of High Intellectual Ability*

Cries about the waste of academic talent, Vernon reminds us, assume that if we were simply to go on a national talent hunt and subsequently send all the high-scoring students to college,

* Reprinted and abridged with the permission of the author and publisher from the article of the same title, *Journal of Educational Research*, 53 (1960), 172-180.

we would avoid a loss to the nation as well as to the individuals concerned. However, Vernon continues, on the basis of studies which correlate high IQ with college success, we can expect that about 40 percent of the students would prove unsuitable (p. 490). The present investigation, although its setting is a high school rather than a college, furnishes evidence to support this expectation. In this New York City high school "gifted" students only are admitted, on the basis of high IQ and entrance examination scores. The subjects of this study had also participated in an enriched junior high school curriculum in which they were identified for high ability and achievement. At the time of the study they were seniors, so that the predictions made in junior high school can be considered long-term—the sort of predictions about which Vernon cautions us (p. 491). The student should recall that the type of "screening" here is typical of that used by many schools and colleges today.

If ability (as indicated by IQ) is not the sole determinant of success, what other factors come into play? As these boys moved into late adolescence what personal and social factors were operating for the success of some and the failure of others? The student should try to piece together the evidence to form two composite pictures, one for each group. How did their environments outside school bolster or disturb their performance in school?

This study was concerned with scholastic underachievement among intellectually superior high school students. The ever-broadening spectrum of our scientific and technological progress, from the harnessing of atomic energy to the conquest of outer space, has placed a special premium on talent and brainpower in all areas of human thought and endeavor. The young people whose scholastic performance lags far behind their intellectual ability represent a serious loss to society in terms of their potential contributions. In addition, failure to achieve at the level of their ability often leads to a depreciation of self-worth accompanied by unhappiness and frustration.

DESIGN OF THE EXPERIMENT

The purpose of this study was to find some answers to the general question, "Why do students of seemingly similar high intellectual ability perform so differently academically?" This investigation proposed to study achieving and underachieving boys of the same high intellectual ability to determine possible causes for the differences in their academic performance. The areas explored for possible significant differences between the two groups were: 1) aptitudes, 2) interests, 3) personal problems, 4) health, 5) home and family background, 6) socio-economic status, 7) reaction to school subjects, 8) reaction to school, 9) out-of-school activities, 10) vocational and college planning and 11) academic performance in junior high school.

Instruments. The following instruments were used for gathering data related to each of these eleven areas—a) Differential Aptitude Tests, b) Kuder Vocational Preference Record, c) Mooney Problem Check List, d) School record, e) a Student Questionnaire of 39 items prepared by the investigator, and f) the Hamburger Scale for rating socio-economic class.

The Experimental Group. The subjects participating in this study were selected from the male population of the senior class of June, 1957 at the Bronx High School of Science in New York City. The experimental group consisted of fifty pairs of boys, each pair composed of an achiever and an underachiever matched on the basis of equivalent I.Q., school entrance examination score, and age.

Definition of terms. Achiever was defined as a student in the top or first quartile of his class with a scholastic average of at least 89 per cent for the tenth and eleventh years. Underachiever was defined as a student in lowest or fourth quartile of the same class with a scholastic average of 79 per cent or less.

Academic Environment of the Study. The study was limited to the Bronx High School of Science because not only was it an ideal source of high ability subjects for this study but also because the investigator has been a teacher and guidance counselor at the school since 1940. The school, which came into being in 1938, was designed to meet the needs of high ability students interested in science and mathematics. The school population is about 2400 one third of

whom are girls. About 800 students are admitted annually, 150 from the elementary schools to the ninth year and 650 from the junior high schools to the tenth year. About 98 per cent of those who are graduated from the school enter college. The school has been described by Wolfe "as being the outstanding exception in this country in providing stimulation and training for bright youngsters with scientific interests."

About four to five times as many students as can be accommodated apply for admission to the school. A program for selecting students has evolved which includes a written examination administered at the school. It is an objective test consisting of two parts: a) English, which includes reading comprehension and vocabulary, and b) Arithmetic. Ninety minutes are allowed for the entire test.

The curriculum of the school aims to prepare students of high ability to meet the admission requirements of liberal arts colleges as well as engineering and technical schools. The subjects are those usually offered in an academic high school. However, the curriculum is enriched by and supplemented with a broad program of elective courses in science and mathematics. In addition, opportunities for acceleration and advanced study are possible by offerings of college-level courses in English, Mathematics, Biology, Chemistry, and Physics.

SELECTION OF THE EXPERIMENTAL GROUP

Criteria for Selection. In matching an achiever with an underachiever, a maximum difference of five points in I.Q., five points in school entrance examination score, and twelve months in age, were used as a basis for obtaining equivalent groups.

The scholastic average which was used to define achiever and underachiever was obtained by calculating the mathematical average of the final grades received in all major subjects during the tenth and eleventh year. Ninth year subjects were not included in the calculations since eighty per cent of the class were in junior high school during that year. Since one of the constant factors in this study was the school environment, marks received in other schools where differences in standards and marking systems might exist, would tend to introduce a variable not included in this study. To obtain equivalent groups optimally matched, only white boys

were selected for this study. In addition, ninth grade entrants were matched only with ninth grade entrants, and those who came from junior high school and entered school in the tenth grade were matched with tenth grade entrants.

Testing Criteria by Preliminary Study. In May, 1956, a preliminary study, involving the class of June, 1956, was undertaken to test the criteria. It was found that forty-two pairs of boys could be matched, using the criteria selected.

Criteria Characteristics of the Experimental Group. In September, 1956, using the criteria established for matching, it was possible to select a maximum of fifty pairs of boys from the class of June, 1957, which had a male population of 468.

The *t*-values for I.Q., entrance examination score, and age indicated that there were no significant differences between the two groups for these criteria. On the other hand, the *t*-value of 23.76 for scholastic average clearly showed that the groups were significantly different in academic achievement, the mean difference being 17.1, with a minimum of 9.5 and a maximum of 37.2.

The entrance examination test score is a composite of the English and the Arithmetic parts. In order to avoid inequalities in matching which might be masked by the composite score, the two groups were compared for the English and Arithmetic scores separately. No significant differences were found between the two groups in English and Arithmetic parts of the entrance examination. The two groups were matched not only for the composite entrance score but also for the English and Arithmetic parts.

FINDINGS

Aptitudes. The results were obtained for fifty pairs of boys on the four Differential Aptitude Tests administered in October, 1956. . . .

The Numerical Ability test discriminated most sharply between the two groups with a highly significant *t*-value of 5.15. The achievers were definitely superior in their ability to understand numerical relationships and in handling numerical concepts.

In addition, the achievers showed definite superiority in Verbal Reasoning, the ability to understand concepts framed in words, the *t*-value being 2.83. Together these two tests are regarded as a measure of general learning ability. The Space Relations and the

Abstract Reasoning tests revealed no significant differences between the groups.

The results of the Scholastic Aptitude Tests for May, 1956 and January, 1957 and the Scholarship Qualifying Tests for April, 1956, which half the subjects had taken, substantiated the DAT findings. The achievers showed significantly greater aptitudes than the under-achievers in the verbal and mathematical areas.

Interests. . . . The interests of the U's were significantly greater in the mechanical area whereas those of the A's were in the scientific, the t-values being significant at the 01 level, 3.18 for the former and 3.01 for the latter. In addition, the A's were more interested in the computational and the U's in the artistic, the t-values being significant at the 05 level. The other six areas showed no significant differences.

Personal Problems. The Mooney Problem Check List, administered in September, 1956, which was used as a measure of the personal problems of the two groups, showed no statistically significant difference in the total number of problems underscored although the A's underscored 723 and the U's 906 problems. . . .

"School" was the only area in which the U's presented significantly more problems than the A's. There were no significant differences in the other six areas.

Of the 210 items of the Mooney, differences in the frequency of responses of the A's and U's significant at and beyond the 05 level occurred for only ten items. Of these, nine were underscored by significantly more U's than A's. Eight of these were concerned with "School" and the other one was "being stubborn." "Family Quarrels" was the only item selected by more A's than U's. In answer to the free question "What problems are troubling you most?" the U's reiterated their concern with school and marks, whereas the A's were interested chiefly in the choice of college and vocation.

Health. The information gathered by the Student Questionnaire regarding health showed no differences between the two groups in weight, height, hearing, speech, general state of health, and physical disabilities. The U's, however, reported significantly more days absent from school for health reasons, the t-values being 2.34. The A's, on the other hand, registered significantly more health complaints, chiefly acne and allergies; the t-value being 2.47.

Family and Home Background. Questionnaire responses indicated no differences between the two groups with respect to a) number of rooms in the home, b) number of people living at home, c) size of family, d) number of disrupted family patterns, and e) birth order of the subjects. Differences were found, however, in the education and occupation of the parents. Using the Edward's Scale for the classification of occupations, more of the fathers of the A's than U's were found in the top three groups—1) professional, 2) semi-professional, and 3) proprietors, managers, and officials. With respect to number of years of schooling completed by parents of the two groups, it was found that the fathers of the A's had significantly more formal education than the mothers, the *t*-value being 2.33. No significant difference between the fathers and mothers of the U's in this respect was found; they had about the same amount of formal education. Significantly more working mothers were reported by the U's than A's, twenty-nine to seventeen, the chi-square value for the difference being 6.43 which is significant at and beyond the 05 level. More of the mothers of the U's than A's were in the lower three occupational groups. Almost two-thirds of them were typists, bookkeepers, secretaries, and saleswomen.

Socio-Economic Status. Using the Hamburger Scale for determining socio-economic level, the A's came from families which were rated significantly higher than those of the U's.

Reaction to School Subjects. In response to the Questionnaire items requiring the groups to name the major school subjects a) liked best, b) found easiest, c) liked least, and d) found most difficult, the following reactions were gleaned: The A's ranked mathematics as the easiest, and mathematics and science as the best liked, English as the most difficult and least liked. The underachievers, on the other hand, selected science as the easiest and best liked subject, and foreign language as the most difficult and least liked. Significantly more A's than U's selected mathematics as the easiest and best liked subject, and English as the most difficult. In contrast, the U's chose science as the easiest and mathematics as the most difficult school subject.

Reaction to School. The criteria selected to measure reaction of the two groups to school were a) attendance, b) deportment, c) participation in extra-curricular activities. These data were obtained

from official school records. According to official attendance records, the U's were absent from school significantly more often than the A's, the t -value being 3.58, significant at the 01 level. School discipline records showed about four times as many offenses recorded against U's than A's, the t -value being 4.07, significant at and beyond the 01 level. With respect to extra-curricular activities, the A's engaged in significantly more, the t -value being 5.35, significant at the 01 level. Twelve U's had records of no extra-curricular activities. The A's engaged significantly more frequently in students government and publications, in science and mathematics clubs, and in the Social Studies club.

Out-of-School and Leisure Time Activities. The U's belonged to significantly more athletic and social clubs and were somewhat more interested in the Scout movement. Insofar as leisure time activities were concerned, the A's tended to spend more time reading whereas the U's were more interested in shop-work activities.

Vocational and College Planning. . . . Significantly more A's planned to enter general fields of science such as mathematics, biology, chemistry and physics. The U's, in contrast, tended more in the direction of applied sciences and technical fields. . . .

Significantly more A's than U's planned to follow a liberal arts college program majoring in science, whereas more U's than A's expected to enter non-science fields such as business administration, accountancy, and the like. Of those who were undecided, . . . more U's than A's selected non-science fields.

The Questionnaire also revealed that more U's than A's, as evidenced by the chi-square value of 4.11, significant at the 05 level, expected their parents to finance completely their college education.

Junior High School Record. The records of forty-three pairs of boys who entered the school in the tenth year from junior high school were studied to determine whether their patterns of academic achievement was developed in the senior high school or had previously been established in the junior high school. It was found that twenty-eight achievers and thirty underachievers had been regarded as intellectually gifted and had been placed in Special Progress (S.P.) classes which had completed the three years of junior high school in two years.

In addition, the results of the reading and arithmetic achievement tests, which had been taken in the ninth year of junior high school,

attested to their superiority. The Stanford Reading Test showed no differences between the groups. Approximately one third of each group made perfect scores and practically the entire group made grade equivalent scores above 10.0, a value derived by extrapolation. The mean grade equivalent score for the entire group was approximately 11 plus. On the other hand, on the New York Arithmetic Test, the equivalent scores of the A's was higher than those of the U's, the *t*-value being 3.12, significant at the 01 level. Nine A's made perfect scores, and practically the entire group scoring above 9.4 also derived by extrapolation. The mean grade equivalent score of the A's was 11.59 and that of the U's was 10.84.

The scholastic record of the two groups for the ninth year indicated that they were performing distinctly differently. Three quarters of the A's attained averages of 89 per cent or better, the lowest being 84. Half of the U's earned less than 80 per cent, fifteen were between 80 and 84, and only four were above 89 per cent. Three quarters of the U's had averages of less than 84 per cent.

CONCLUSIONS

1. *Aptitudes.* Although the pairs were matched for equivalent I.Q. and school entrance examination, the achievers proved to be distinctly superior to the underachievers in mathematical and verbal aptitudes, particularly in the former.

2. *Interests.* The interest patterns of the two groups were distinctly different. The interests of the achievers were greater in mathematics and science whereas those of the underachievers were in the mechanical and artistic areas.

3. *Personal Problems.* While the chief concern of the underachiever appeared to be his present scholastic inadequacies, the achiever's was primarily thinking about the future, college and vocational choices.

4. *Health.* Although the U's reported twice as many days absent from school for health reasons, and official attendance records disclosed that the U's were absent from school significantly more frequently, they registered fewer specific health complaints on the Student Questionnaire, and underscored fewer items in the Health and Physical Development area of the Mooney. No evidence was found to lead one to believe that the two groups differed signifi-

cantly in health. It seemed likely that the more frequent absence from school for health reasons reported by the U's might not necessarily have been the result of physical illness.

5. *Home and Family Background.* Although the physical aspects of the families of the two groups were very much alike, significant differences in the education and occupation of the parents existed. More of the fathers of the A's were in the top three occupational groups and they had more formal schooling than their wives. More of the mothers of the U's were working and they had at least as much schooling as their husbands.

6. *Socio-Economic Status.* As expected, the families of the A's were rated higher on the Hamburger Socio-Economic scale.

7. *Reaction to School Subjects.* The selection by the achievers of mathematics as the easiest, and science and mathematics as the best liked school subjects, was probably a reflection of their superior aptitude and greater interest in these areas. Similarly, the distaste of the U's for mathematics mirrored the difficulty which they had with this subject.

The negative reactions of the A's to English might be an expression of their science-mathematics preference. The selection of science as the easiest and best liked subject by the U's might be explained by the fact that the sciences with their concomitant laboratory opportunities offered an outlet for their mechanical interests.

8. *Reaction to School.* It was not surprising to find that the U's evidenced negative attitudes toward school, the major source of their personal problems, in terms of poorer attendance records, more recorded disciplinary offenses, and participation in fewer extra-curricular activities. In general, the U's were more recalcitrant, less conforming, and less happy at school. The achievers, in contrast, were more conforming, rarely broke school regulations, participated in more school activities and assumed positions of leadership and responsibility.

9. *Out-of-School and Leisure Activities.* The greater participation of the U's in out-of-school organizations such as Social and Athletic clubs, and the Scouts, may have been a substitute for school activities.

10. *Vocational and College Planning.* The achievers appeared to regard college as preparatory for a career in science with the ex-

pectation of going on to graduate school for specialization. The underachievers tended to think of college in terms of direct vocational preparation; those going into the sciences were planning careers in applied and technical fields. However, a substantial number of U's planned to prepare for and enter non-science fields.

11. *Junior High School Academic Performance.* Notwithstanding the superior intellectual ability of the two groups, the ninth year junior high school record left little doubt that the two groups performed differently in terms of academic achievement. In general, the achievers maintained their high scholastic record while the performance of the underachievers deteriorated. The difference in the mean scholastic average of the two groups was twice as great in high school as in junior high school. It appeared probable that the factors relating to scholastic underachievement of this group may have been operating before these students entered the high schools.

LEONARD SCHATZMAN and ANSELM STRAUSS
Coe College, University of Chicago

Social Class and Modes of Communication *

Differences in the social class origin of students are most readily observable in the urban school. A middle-class teacher with a world of perceptions, concepts, attitudes, and skills quite unlike the world of the low-class children she struggles to teach could supply drama and pathos for many novels. In fact, if the reader substitutes for the middle-class interviewer in the following report the middle-class teacher, he will perceive vividly some aspects of the school problem. The most severe deprivation that lower-class children suffer is not exclusively material poverty, but the impoverishment of an environment that is im-

* Reprinted and abridged with the permission of the University of Chicago Press from the article of the same title, *American Journal of Sociology*, 60 (1955), 329-338.

mediate, concrete, and undiscriminated. Children frequently live in such environments until the age of six. Vernon questions whether it is possible for the school to overcome the effects of such early deprivation (p. 491). Burt's view of the genetic determination of intelligence and the assortative mating of individuals of high intelligence does not make the picture less pessimistic. It would seem clear that the education of lower-class children must always be restricted until social and economic progress changes their preschool environment.

The lack of verbal ability in lower-class children is often claimed to be the cause of their poor performance on IQ tests. Examination of the evidence in this report indicates that the difficulty may be far deeper and connected with cognitive learning. In investigating the relationship between communication and cognition the student should answer these questions: (1) What differences were there between the way lower-class and middle-class individuals organized their thoughts? (2) How does the "buck-shot" listing of images and incidents which characterized lower-class communication reflect what Bruner has described about the two major modes of thinking (pp. 258-259)? (3) Why did lower-class individuals frequently fail to explain behavior in terms of motives?

The lack of verbal ability in lower-class individuals can also be seen as a communication problem. Lorge has defined redundancy, context, and sensitivity to audience (pp. 326-338). How do middle-class and lower-class communication differ in these respects?

Perhaps the crucial question is: In teaching concepts and in communicating with lower-class children, what guidelines does the socio-psychological research reported here furnish the middle-class teacher?

Common assumptions suggest that there may be important differences in the thought and communication of social classes. Men live in an environment which is mediated through symbols. By naming, identifying, and classifying, the world's objects and events are

perceived and handled. Order is imposed through conceptual organization, and this organization embodies not just anybody's rules but the grammatical, logical, and communicative canons of groups. Communication proceeds in terms of social requirements for comprehension, and so does "inner conversation" or thought. Both reasoning and speech meet requirements of criticism, judgment, appreciation, and control. Communication across group boundaries runs the danger—aside from sheer language difficulties—of being blocked by differential rules for the ordering of speech and thought.

If these assumptions are correct, it follows that there should be observable differences in communication according to social class and that these differences should not be merely matters of degree of preciseness, elaboration, vocabulary, and literary style. It follows also that the modes of thought should be revealed by modes of speaking.

Our data are the interview protocols gathered from participants in a disaster. The documents, transcribed from tape, contain a wealth of local speech. Respondents had been given a relatively free hand in reporting their experiences, and the interviews averaged twenty-nine pages. These seemed admirably suited to a study of differences between social classes in modes of communication and in the organization of perception and thought. We used them also to explore the hypothesis that substantial intraclass differences in the organization of stories and accounts existed; hence low-class respondents might fail to satisfy the interviewer's canons of communication.

Approximately 340 interviews were available, representing random sampling of several communities ravaged by a tornado. Cases were selected by extreme position on educational and income continuums. Interviewees were designated as "lower" if education did not go beyond grammar school and if the annual family income was less than two thousand dollars. The "upper" group consisted of persons with one or more years of college education and annual incomes in excess of four thousand dollars. These extremes were purposely chosen for maximum socioeconomic contrast and because it seemed probable that nothing beyond formal or ritual communication would occur between these groups.

Cases were further limited by the following criteria: age (twenty-one to sixty-five years), race (white only), residence (native of Arkansas and more than three years in the community), proximity (either in the disaster area or close by) good co-operation in interview (as rated by interviewer), and less than eight probes per page (to avoid a rigid question-answer style with consequent structuring of interview by the interviewer's questions). The use of these criteria yielded ten upper-group cases, which were then matched randomly with ten from the lower group.

Differences between Classes

Differences between the lower and upper groups were striking; and, once the nature of the difference was grasped, it was astonishing how quickly a characteristic organization of communication could be detected and described from a reading of even a few paragraphs of an interview. The difference is not simply the failure or success—of lower and upper groups, respectively—in communicating clearly and in sufficient detail for the interviewer's purposes. Nor does the difference merely involve correctness or elaborateness of grammar or use of a more precise or colorful vocabulary. The difference is a considerable disparity in (a) the number and kinds of perspectives utilized in communication; (b) the ability to take the listener's role; (c) the handling of classifications; and (d) the frameworks and stylistic devices which order and implement the communication.

Perspective or Centering

By perspective or centering is meant the standpoint from which a description is made. Perspectives may vary in number and scope. The flexibility with which one shifts from perspective to perspective during communication may vary also.

Lower Class. Almost without exception any description offered by a lower-class respondent is a description as seen through his *own* eyes; he offers his own perceptions and images directly to the listener. His best performance is a straight, direct narrative of events

as he saw and experienced them. He often locates himself clearly in time and place and indicates by various connective devices a rough progression of events in relation to his activities. But the developmental progression is only in relation to himself. Other persons and their acts come into his narrative more or less as he encountered them. In the clearest interviews other actors are given specific spatial and temporal location, and sometimes the relationships among them or between them and himself are clearly designated.

The speaker's images vary considerably in clarity but are always his own. Although he may occasionally repeat the stories of other persons, he does not tell the story as though he were the other person reconstructing events and feelings. He may describe another person's act and the motive for it, with regard to himself, but this is the extent of his role-taking—he does not assume the role of another toward still others, except occasionally in an implicit fashion: "Some people was helping other people who was hurt." This limitation is especially pronounced when the behavior of more than two or three persons is being described and related. Here the description becomes confused: At best the speaker reports some reactions, but no clear picture of interaction emerges. The interaction either is not noticed or is implicitly present in the communication ("We run over there to see about them, and they was alright"). Even with careful probing the situation is not clarified much further. The most unintelligible speakers thoroughly confound the interviewer who tries to follow images, acts, person, and events which seem to come out of nowhere and disappear without warning.

Middle Class. The middle class can equal the best performance of the lower class in communicating and elaborating a direct description. However, description is not confined to so narrow a perspective. It may be given from any of several standpoints: for instance, another person, a class of persons, an organization, an organizational role, even the whole town. The middle-class speaker may describe the behavior of others, including classes of others, from their standpoints rather than from his, and he may include sequences of acts as others saw them. Even descriptions of the speaker's own behavior often are portrayed from other points of view.

Correspondence of Imagery between Speaker and Listener

Individuals vary in their ability to see the necessity for mediating linguistically between their own imagery and that of their listeners. The speaker must know the limits within which he may assume a correspondence of imagery. When the context of the item under discussion is in physical view of both, or is shared because of similarity of past experience, or is implicitly present by virtue of a history of former interaction, the problem of context is largely solved. But when the context is neither so provided nor offered by the speaker, the listener is confronted with knotty problems of interpretation. In the accounts of the most unintelligible respondents we found dream-like sets of images with few connective, qualifying, explanatory, or other context-providing devices. Thus, the interviewer was hard pressed to make sense of the account and was forced to probe at every turn lest the speaker figuratively run away with the situation. The respondents were willing and often eager to tell their stories, but intention to communicate does not always bring about clear communication. The latter involves, among other requirements, an ability to hear one's words as others hear them.

Lower Class. Lower-class persons displayed a relative insensitivity to disparities in perspective. At best, the respondent corrected himself on the exact time at which he performed an act or became aware that his listener was not present at the scene and so located objects and events for him. On occasion he reached a state of other-consciousness: "You can't imagine if you wasn't there what it was like." However, his assumption of a correspondence in imagery is notable. There is much surnaming of persons without genuine identification, and often terms like "we" and "they" are used without clear referents. The speaker seldom anticipates responses to his communication and seems to feel little need to explain particular features of his account. He seldom qualifies an utterance, presumably because he takes for granted that his perceptions represent reality and are shared by all who were present. Since he is apt to take so much for granted, his narrative lacks depth and richness and contains almost no qualifications and few genuine illustrations. The hearer very often is confronted with a descriptive fragment that supposedly

represents a more complete story. The speaker may then add phrases like "and stuff like that" or "and everything." Such phrasing is not genuine summation but a substitute for detail and abstraction. Summary statements are virtually absent, since they signify that speakers are sensitive to the needs of listeners. Certain phrases that appear to be summaries—such as "That's all I know" and "That's the way it was"—merely indicate that the speaker's knowledge is exhausted. Other summary-like phraseologies, like "It was pitiful," appear to be asides, reflective of self-feeling or emotion rather than résumés of preceding detail.

Middle Class. The middle-class respondent also makes certain assumptions about the correspondence of the other's images with his own. Nevertheless, in contrast with the lower group, he recognizes much more fully that imagery may be diverse and that context must be provided. Hence he uses many devices to supply context and to clarify meaning. He qualifies, summarizes, and sets the stage with rich introductory material, expands themes, frequently illustrates, anticipates disbelief, meticulously locates and identifies places and persons—all with great complexity of detail. He depends less on saying "You know"; he insists upon explaining if he realizes that a point lacks plausibility or force. Hence he rarely fails to locate an image, or series of images, in time or place. Frequent use of qualifications is especially noteworthy. This indicates not only multiple centering but a very great sensitivity to listeners, actual and potential—including the speaker himself.

In short, the middle-class respondent has what might be called "communication control," at least in such a semiformal situation as the interview. Figuratively, he stands between his own images and the hearer and says, "Let me introduce you to what I saw and know." It is as though he were directing a movie, having at his command several cameras focused at different perspectives, shooting and carefully controlling the effect. By contrast, the lower-class respondent seems himself more like a single camera which unreels the scene to the audience. In the very telling of his story he is more apt to lose himself in his imagery. The middle-class person—by virtue, we would presume, of his greater sensitivity to his listener—stands more outside his experience. He does not so much tell you what he saw as fashion a story about what he saw. The story may be accurate in varying degrees, although, in so far as it is an organized account, it

has both the virtues and the defects of organization. The comparative accuracies of middle- and lower-class accounts are not relevant here; the greater objectivity of the former merely reflects greater distance between narrator and event.

In organizing his account, the middle-class respondent displays parallel consciousness of the other and himself. He can stop mid-stream, take another direction, and, in general, exert great control over the course of his communication. The lower-class respondent seems to have much less foresight, appearing to control only how much he will say to the interviewer, or whether he will say it at all, although presumably he must have some stylistic controls not readily observable by a middle-class reader.

Classifications and Classificatory Relations

Lower Class. Respondents make reference mainly to the acts and persons of particular people, often designating them by proper or family names. This makes for fairly clear denotation and description, but only as long as the account is confined to the experiences of specific individuals. There comes a point when the interviewer wishes to obtain information about classes of persons and entire organizations as well as how they impinged upon the respondent, and here the lower-class respondent becomes relatively or even wholly inarticulate. At worst he cannot talk about categories of people or acts because, apparently, he does not think readily in terms of classes. Questions about organizations, such as the Red Cross, are converted into concrete terms, and he talks about the Red Cross "helping people" and "people helping other people" with no more than the crudest awareness of how organizational activities interlock. At most the respondent categorizes only in a rudimentary fashion: "Some people were running; other people were looking in the houses." The interviewer receives a sketchy and impressionistic picture. Some idea is conveyed of the confusion that followed upon the tornado, but the organizing of description is very poor. The respondent may mention classes in contrasting juxtaposition (rich and poor, hurt and not-hurt), or list groups of easily perceived, contrasting actions, but he does not otherwise spell out relations between these classes. Neither does he describe a scene systematically

in terms of classes that are explicitly or clearly related, a performance which would involve a shifting of viewpoint.

It is apparent that the speakers think mainly in particularistic or concrete terms. Certainly classificatory thought must exist among many or all the respondents; but, in communicating to the interviewer, class terms are rudimentary or absent and class relations implicit: relationships are not spelled out or are left vague. Genuine illustrations are almost totally lacking, either because these require classifications or because we—as middle-class observers—do not recognize that certain details are meant to imply classes.

Middle Class. Middle-class speech is richly interlarded with classificatory terms, especially when the narrator is talking about what he saw rather than about himself. Typically, when he describes what other persons are doing, he classifies actions and persons and more often than not explicitly relates class to class. Often his descriptions are artistically organized around what various categories of persons were doing or experiencing. When an illustration is offered, it is clear that the speaker means it to stand for a general category. Relief and other civic organizations are conceived as sets or classes of co-ordinated roles and actions; some persons couch their whole account of the disaster events in organizational terms, hardly deigning to give proper names or personal accounts. In short, concrete imagery in middle-class communication is dwarfed or overshadowed by the prevalence and richness of conceptual terminology. Organization of speech around classifications comes readily, and undoubtedly the speaker is barely conscious of it. It is part and parcel of his formal and informal education. This is not to claim that middle-class persons always think with and use classificatory terms, for doubtless this is not true. Indeed, it may be that the interview extracts from them highly conceptualized descriptions. Nonetheless, we conclude that, in general, the thought and speech of middle-class persons is less concrete than that of the lower group.

Organizing Frameworks and Stylistic Devices

One of the requirements of communication is that utterances be organized. The principle of organization need not be stated explicitly by the speaker or recognized by the listener. Organizing

frames can be of various sorts. Thus an ordering of the respondents' description is often set by the interviewer's question, or the speaker may set his own framework ("There is one thing you should know about this"). The frame can be established jointly by both interviewer and respondent, as when the former asks an open-ended question within whose very broad limits the respondent orders his description in ways that strike him as appropriate or interesting. The respondent, indeed, may organize his account much as though he were telling a special kind of story or drama, using the interviewer's questions as hardly more than general cues to what is required. The great number of events, incidents, and images which must be conveyed to the listener may be handled haphazardly, neatly, dramatically, or sequentially; but, if they are to be communicated at all, they must be ordered somehow. Stylistic devices accompany and implement these organizing frames, and the lower and upper groups use them in somewhat different ways.

Lower Class. The interviewer's opening question, "Tell me your story of the tornado," invites the respondent to play an active role in organizing his account; and this he sometimes does. However, with the exception of one person who gave a headlong personal narrative, the respondents did not give long, well-organized, or tightly knit pictures of what happened to them during and after the tornado. This kind of general depiction either did not occur to them or did not strike them as appropriate.

The frames utilized are more segmental or limited in scope than those used by the middle class. They appear to be of several kinds and their centering is personal. One is the personal narrative, with events, acts, images, persons, and places receiving sequential ordering. Stylistic devices further this kind of organization: for instance, crude temporal connectives like "then," "and," and "so" and the reporting of images or events as they are recollected or as they appear in the narrative progression. Asides may specify relationships of kinship or the individuals' location in space. But, unless the line of narrative is compelling to the speaker, he is likely to wander off into detail about a particular incident, where the incident in turn then provides a framework for mentioning further events. Likewise, when a question from the interviewer breaks into the narrative, it may set the stage for an answer composed of a number of images or an incident. Often one incident becomes the trigger for another,

and, although some logical or temporal connection between them may exist for the speaker, this can scarcely be perceived by the interviewer. Hence the respondent is likely to move out of frames quickly. The great danger of probes and requests for elaboration is that the speaker will get far away from the life-line of his narrative—and frequently far away from the interviewer's question. As recompense the interviewer may garner useful and unexpectedly rich information from the digressions, although often he needs to probe this material further to bring it into context. General questions are especially likely to divert the speaker, since they suggest only loose frames; or he may answer in general, diffuse, or blurred terms which assume either that the listener was there too or that he will put meaningful content into the words. If a question is asked that concerns abstract classes or is "above" the respondent—a query, say, about relief organizations—then very general answers or concrete listing of images or triggering of images are especially noticeable. When the interviewer probes in an effort to get some elaboration of an occurrence or an expansion of idea, he commonly meets with little more than repetition or with a kind of "buckshot" listing of images or incidents which is supposed to fill out the desired picture. The lack of much genuine elaboration is probably related to the inability to report from multiple perspectives.

One requirement of the interview is that it yield a fairly comprehensive account of the respondent's actions and perceptions. With the lower-class respondent the interviewer, as a rule, must work very hard at building a comprehensive frame directly into the interview. This he does by forcing many subframes upon the respondent. He asks many questions about exact time sequence, placement and identification of persons, expansion of detail, and the like. Especially must he ask pointed questions about the relations of various personages appearing in the account. Left to his own devices, the respondent may give a fairly straight-forward narrative or competently reconstruct incidents that seem only partially connected with each other or with his narrative. But the respondent seldom voluntarily gives both linear and cross-sectional pictures.

The devices used to implement communication are rather difficult to isolate, perhaps because we are middle class ourselves. Among the devices most readily observable are the use of crude chronological notations (e.g., "then, . . . and then"), the juxtaposing or direct

contrasting of classes (e.g., rich and poor), and the serial locating of events. But the elaborate devices that characterize middle-class interviews are strikingly absent.

Middle Class. Without exception middle-class respondents imposed over-all frames of their own upon the entire interview. Although very sensitive generally to the needs of the interviewer, they made the account their own. This is evidenced sometimes from the very outset; many respondents give a lengthy picture in answer to the interviewer's invitation, "Tell me your story." The organizing frame may yield a fluid narrative that engulfs self and others in dense detail; it may give a relatively static but rich picture of a community in distress; or, by dramatic and stage-setting devices, it may show a complicated web of relationships in dramatic motion. The entire town may be taken as the frame of reference and its story portrayed in time and space.

Besides the master-frame, the middle-class respondent utilizes many subsidiary frames. Like the lower-class person, he may take off from a question. But, in doing so—especially where the question gives latitude by its generality or abstractness—he is likely to give an answer organized around a subframe which orders his selection and arrangement of items. He may even shift from one image to another, but rarely are these left unrelated to the question which initially provoked them. He is much more likely also to elaborate than to repeat or merely to give a scattered series of percepts.

One prerequisite for the elaboration of a theme is an ability to depart from it while yet holding it in mind. Because he incorporates multiple perspectives, the respondent can add long asides, discuss the parallel acts of other persons in relation to himself, make varied comparisons for the enrichment of detail and comprehension—and then can return to the original point and proceed from there. Often he does this after first preparing his listener for the departure and concludes the circuit with a summary statement or a transitional phrase like "well—anyhow" that marks the end of the digression.

The stylistic devices utilized by any respondent are many and varied. But each speaker uses some devices more frequently than others, since certain ones are more or less appropriate to given frames. There is no point in spelling out the whole range of devices; they are of the sort used in any clear detailed narrative and effective exposition. If the respondent is pressed to the limit of his

ability in explaining a complex point or describing a complicated scene, he calls into play resources that are of immensely high order. Sometimes a seemingly simple device will turn out on closer inspection to demand a sophisticated handling of communication—for instance, the frequent and orderly asides that break into exposition or narrative and serve with great economy to add pertinent detail.

Discussion

Only if the situation in which the respondent spoke is carefully taken into account will we be on safe ground in interpreting class differences. Consider, first, the probable meaning of the interview for the middle-class respondents. Although the interviewer is a stranger, an outsider, he is a well-spoken, educated person. He is seeking information on behalf of some organization, hence his questioning not only has sanction but sets the stage for both a certain freedom of speech and an obligation to give fairly full information. The respondent may never before have been interviewed by a research organization, but he has often talked lengthily, fairly freely, and responsibly to organizational representatives. At the very least he has had some experience in talking to educated strangers. We may also suppose that the middle-class style of living often compels him to be very careful not to be misunderstood. So he becomes relatively sensitive to communication *per se* and to communication with others who may not exactly share his viewpoints or frames of reference.

Communication with such an audience requires alertness, no less to the meanings of one's own speech than to the possible intent of the other's. Role-taking may be inaccurate, often, but it is markedly active. Assessing and anticipating reactions to what he has said or is about to say, the individual develops flexible and ingenious ways of correcting, qualifying, making more plausible, explaining, rephrasing—in short, he assumes multiple perspectives and communicates in terms of them. A variety of perspectives implies a variety of ways of ordering or framing detail. Moreover, he is able to classify and to relate classes explicitly, which is but another way of saying that he is educated to assume multiple perspectives of rather wide scope.

It would certainly be too much to claim that middle-class persons always react so sensitively. Communication is often routinized, and much of it transpires between and among those who know each other so well or share so much in common that they need not be subtle. Nor is sensitive role-taking called forth in so-called "expressive behavior," as when hurling invective or yelling during a ball game. With the proviso that much middle-class speech is uttered under such conditions, it seems safe enough to say that people of this stratum can, if required, handle the more complex and consciously organized discourse. In addition to skill and perspicacity, this kind of discourse requires a person who can subtly keep a listener at a distance while yet keeping him in some degree informed.

Consider now, even at risk of overstating the case, how the interview appears to the lower group. The interviewer is of higher social class than the respondent, so that the interview is a "conversation between the classes." It is entirely probable that more effort and ability are demanded by cross-class conversation of this sort than between middle-class respondent and middle-class interviewer. It is not surprising that the interviewer is often baffled and that the respondent frequently misinterprets what is wanted. But misunderstanding and misinterpretation are only part of the story.

Cross-class communication, while not rare, probably is fairly formalized or routinized. The communicants know the ritual steps by heart, and can assume much in the way of supporting context for phrase and gesture. The lower-class person in these Arkansas towns infrequently meets a middle-class person in a situation anything like the interview. Here he must talk at great length to a stranger about personal experiences, as well as recall for his listener a tremendous number of details. Presumably he is accustomed to talking about such matters and in such detail only to listeners with whom he shares a great deal of experience and symbolism, so that he need not be very self-conscious about communicative technique. He can, as a rule, safely assume that words, phrases, and gestures are assigned approximately similar meanings by his listeners. But this is not so in the interview or, indeed, in any situation where class converses with class in nontraditional modes.

There still remains the question of whether the descriptions of perceptions and experiences given by the lower-class respondent are

merely inadequate or whether this is the way he truly saw and experienced. Does his speech accurately reflect customary "concrete" modes of thought and perception, or is it that he perceives in abstract and classificatory terms, and from multiple perspectives, but is unable to convey his perceptions? Unless one assumes that, when talking in familiar vein to familiar audiences, speech and gesture incorporate multiple perspectives, which is, as we have already indicated, improbable, one concludes that speech does in some sense reflect thought. The reader is perhaps best left at this point to draw his own conclusions, although we shall press upon him certain additional evidence and interpretation arising from examination of the interviews.

In any situation calling for a description of human activities it is necessary to utilize motivational terminology, either explicitly or implicitly, in the very namings of acts. In the speech of those who recognize few disparities of imagery between themselves and their listeners, explicit motivational terms are sparse. The frequent use among the lower class of the expression "of course" followed by something like "They went up to see about their folks" implies that it is almost needless to say what "they" did, much less to give the reason for the act. The motive ("to see about") is implicit and terminal, requiring neither elaboration nor explanation. Where motives are explicit ("They was needin' help, so we went on up there"), they are often gratuitous and could just as well have been omitted. All this is related to preceding discussions of single centering and assumed correspondence of imagery. To the speaker it was quite clear why people did what they did. There was no need to question or to elaborate on the grounds for acts. Under probing the respondent did very little better: he used motivational terms but within a quite narrow range. The terms he used ordinarily reflected kinship obligations, concern for property, humanitarian ("help") sentiments, and action from motives of curiosity ("We went down to see"). Such a phrase as "I suppose I went to her house because I wanted reassurance" would rarely occur.

Middle-class persons exhibit familiarity with a host of distinct "reasons" for performing particular acts. Their richness in thinking allows activities to be defined and described in a great variety of ways. Here, indeed, is an instrument for breaking down diffuse

images ("They was runnin' all over") into classes of acts and events. The middle-class person is able to do this, for one thing, because he possesses an abstract motivational terminology. Then, too, the fine and subtle distinctions for rationalizing behavior require devices for insuring that they will be grasped by the hearer. In a real sense the need to explain behavior can be linked with the need to communicate well—to give a rational account as well as to be objective. Hence, there is a constant flow of qualifying and generalizing terms linked with motivational phraseology ("I don't know why, but it could be he felt there was no alternative . . .").

It is not surprising to find the middle class as familiar with elements of social structure as with individual behavior. Assuredly, this familiarity rests not only upon contact with institutions but upon the capacity to perceive and talk about abstract classes of acts. The lower-class person, on the other hand, appears to have only rudimentary notions of organizational structure—at least of relief and emergency agencies. Extended contact with representatives of them, no doubt, would familiarize him not only with organizations but with thinking in organizational, or abstract, terms. The propensity of the lower class to state concretely the activities of relief organizations corroborates the observation of Warner that the lowest strata have little knowledge or "feel" for the social structures of their communities.¹ It also suggests the difficulty of conveying to them relatively abstract information through formal media of communication.

It may be that rural townspeople of the lower class are not typical of the national or urban low strata. This raises the question—vital to urban sociology but to which currently there is no adequate answer—of whether pockets of rural-minded folk cannot live encapsulated in the city² and, indeed, whether lower-class persons have much opportunity to absorb middle-class culture without themselves beginning the route upward, those remaining behind remaining less urban.

¹ W. L. Warner, *American Life: Dream and Reality* (Chicago: University of Chicago Press, 1953), pp. 193–94.

² David Riesman, "Urbanity and the Urban Personality," in *Proceedings of the Fourth Annual Symposium, The Human Development Bulletin* (Chicago: University of Chicago, 1953), p. 37.

[CHAPTER 9]

Social Learning: Isn't It Always Better in a Group?

Introduction

Group methods of instruction and organization may be psychologically sound ways of achieving some educational objectives. In a democratic society we do not want students to develop excessive dependence on authority figures; we want them to acquire the skills necessary for working harmoniously and productively with their peers, and we want them to develop leadership which maintains some degree of sensitivity to the needs and desires of the membership. However, when group methods of instruction become the exclusive means employed, irrespective of the particular educational task (whether it be learning to spell or learning to cooperate), and when such prohibition of nongroup methods is strongly defended as the only effective and democratic means of teaching, it is hard to resist the clinical judgment that some irrational fear of authority has seized the staunch defenders of the group.

The frankly managerial role of the teacher in the educative process has been assumed throughout this book. It has been fully assumed by Skinner when he indicates what guidance is desirable (and even by Bruner in the discussion of learning as discovery). Furthermore, the teacher cannot pursue educational objectives entirely of his own choosing. Nor, certainly, can the student. Ultimately, the schools, and the teachers and students in them, must conform to the demands of the society which builds and supports them. This does not mean that objectives cannot be adjusted to

the local requirements of particular schools, teachers, and students. It does mean that society expects the school to get certain jobs done, and society, at times, can be unequivocal about what these are.

This collection of articles and the accompanying editorial comments suggest a psychological framework within which teachers can make decisions. What is required is that group or social learning become the object of carefully planned research. The reports which follow either review the problems of such research (and there are many), or report on how research has been carried out. They all suggest that relatively informal assessment of student opinion on how well they liked or disliked a particular method of instruction is hardly enough. We have seen in the study by Siegel and Macomber on large-group and televised instruction that how well the students liked or did not like the particular class setting was not reflected in their examination scores (pp. 409-414). McKeachie provides more evidence on this point (pp. 554-556). One confuses theory of motivation with theory of learning if one equates all happy classroom experiences with profitable ones.

In the popular press there have been many indications of a shift in the public viewpoint concerning appropriate objectives for the school. The trend of opinion seems to favor the development of fairly well defined cognitive abilities and skills which the student will find necessary and useful in his advanced education or vocationally useful in his postgraduate job. This has meant that the broad attitudinal objectives of the period of depression and reform which preceded World War II are no longer emphasized. Parents, for example, are now more specifically concerned with improvement in reading skills, skills in calculation, educational advancement, and favorable job placement. This does not mean that their interest in the school's development of leadership, good group membership, satisfactory interpersonal relations, and even sturdy character, has vanished. But in the competition with the more tangible and specific academic goals, the attitudinal objectives have become less important.

If this is occurring, it may well indicate that group methods of instruction in many instances will become less useful than those methods which can better guarantee the educational outcomes now

avored. Group instruction was often the vehicle of attitude change. Specifically, as Lorge points out, the groups were frequently experiments in democratic living which later, unfortunately, became rather institutionalized. Undoubtedly this is one reason why group instruction is so vehemently defended against the critic, even when he attempts only to specify its limited usefulness. To question their method, for the group stalwarts, is to attack their objective, which they see as nothing less than the promotion of democracy and democratic living. In the future, it may be necessary frequently to point out that rejection or limited acceptance of group methods of instruction is not a rejection of democracy.

Relationship of Readings in Chapter 9

The proper type of research can help us identify the important variables and beneficial effects of social learning. Lorge, in the first article, describes what the major characteristics of such research ought to be. Marquart's investigation of group problem solving illustrates the type of research Lorge refers to. McKeachie continues Lorge's discussion of research; he shows how muddled is the issue of student-centered versus instructor-centered procedures and how the research on this question should be constructed. Finally, the article by Zander and Cohen reports an investigation of the group as an interaction process and thereby suggests new variables important in social learning.

IRVING LORGE

Late Professor of Education

Teachers College, Columbia University

Groupness of the Group*

Are individuals more successful in solving problems when they work alone or when they work in a group? We are reminded by Lorge that before we can answer the question we must have some definition of what a group is. In the Introduction (p. 527) the group was referred to as a method of instruction which can help students achieve particular educational goals. For example, a reading group in the first grade may be a means of giving common instruction to pupils who possess the same level of skill or have the same difficulties in reading. Such groups, in Lorge's terminology, lack "groupness." Psychologically, the group is discussed in terms of the various leader and member roles which are enacted in it and the impact of this behavior on each of the members. The group interaction described by Zander and Cohen in the article which follows later (pp. 559-565) exemplifies some aspects of this groupness. The group which has this quality of groupness may still be an instructional group, but its objectives would relate to social organization, the development of cohesiveness, skill in social interaction, etc.

Lorge's distinction between *ad hoc* (experimental) groups and dynamic (on-the-job) groups recalls the difficulties, referred to by Melton and others, which occur in translating laboratory findings for application in practical situations. The situational factors may actually impose greater restrictions on generalizations in the case of groups than individuals. For example, what limitations does this translation place on generalizations which can be made about individual versus group

* Reprinted and abridged with the permission of Mrs. Irving Lorge and the Abraham Magazine Service, from the article of the same title, *Journal of Educational Psychology*, 46 (1955), 449-456.

superiority in learning? What concrete proposals for research on groups in the real school situation can the student suggest (in the light of Lorge's discussion)? How would you test for transfer of group learning of social skills?

There is a tendency, among educational psychologists, today to accept some objectives, procedures and content as legitimate curricular goals without assaying the attainability of them. For instance, there seems to be considerable emphasis in the educational and psychological literature on the attainment of groupness and group dynamics. There can be little quarrel with the stipulation of such goals provided that some attempts be made to find out whether the goals as ends, or as means to other ends, are realized or realizable.

From a brief review of group dynamics, it is evident that much attention has been given to the purpose of group dynamics at the elementary school level. The expressed purposes range from the mere learning of specific skills and knowledges to the attainment of the broader aspects of democratic living. For instance, Thelen believes that instructional efficiency can be increased through the teaching of group dynamics in the classroom. He suggests that all students in a subgroup can be brought up to the level originally possessed by the most skilled person in it, if group methods of teaching are intelligently used. It is true that Thelen feels that a primary purpose of group methods in teaching and learning is in the promotion of creative thinking and social organization. When the class discusses and settles the question of how work shall be organized and individual efforts co-ordinated, he feels that the individuals and the class gain in mastery of decision-making and in problem-solving. Learning in groups is said to satisfy individual needs so that energies may be devoted to solving achievement problems. Indeed, Thelen tends to emphasize not only the individual's gains in the mastering of content, in the accepting of self responsibility, in the motivating of learning, but also stresses the acquiring of group process goals in group feeling and in the understanding of democratic adjustments.

With the current interest in the group and in group dynamics,

however, little has been said about what makes a group *a group* as distinguished either from an aggregation or, indeed, from an individual. The popular preference for the word "group" suggests a nuance of an organic unit or a cohesive whole. In the literature, however, the structure of the group is ill-defined and ambiguous. In the references that have been reviewed, the structure of the group, as used in elementary school teaching, may be broadly dichotomized, (1) as that in which the group is considered more in terms of method of instruction than in terms of its organization; and (2) as that in which the group is referred to in terms of rôle and interaction patterns.

The first obviously does not define a group at all, but gives the flavor of the classroom climate in relation to relative limits placed on child activities both by teacher demands and decisions as well as by the environmental constraints and limitations considered necessary for teaching and learning. Essentially, group structure merely—or really suggests the delicate balancing between the Charybdis of permissiveness and the Scylla of control. The second category asserts that there are differences between an aggregate and a group such as the fact that a group's members have common needs, interests and purposes. This, too, fails fully to specify a group, for in the schoolroom in so far as members have different needs, interests and purposes, they would form different groups. Perhaps, the closest indication of the differentia is in the process of interaction among the members of a group. Such an approach to the group leads to difficulties in the appraisal of and even the description of the communications and transactions among the members. Yet, the groupness of any group must represent some sharing of experiences, knowledges, feelings and purposes. Perhaps groupness, or as one writer terms it, "we-ness" is in the sharing of resources, be they intellectual, emotional or material, to accomplish a group purpose or several group purposes.

Even now, however, there is little or no evidence that suggests the degree to which groups achieve any one group purpose, or the extent to which the members as individuals gain from group experience in such objectives as social skills and knowledge, or, more significantly, in the ability to transfer such social gains to other groups. It is too early to expect that the effect of the group experi-

ence upon the school children can be appraised. At least, for research, there are hypotheses that should be tested such as: (a) any group benefits are shared equally by all individual members; (b) individual learning is maximized through the group; (c) group experiences strengthen the individual's need to learn; (d) group experience allows the individual to develop at his own rate to maturity; and so on.

It is my belief, however, that it will be very difficult to design research in the effect of the groupness of the group upon the behavior of its individual members largely because the group is so ill-defined.

In the course of summarizing the literature on experimentation with groups in problem solving, we became aware that the psychologist used the term "group" in at least five different meanings. This is not strange. For the lexicographer considers a group to be a "number of persons (or things) gathered together and, thereby, forming a recognizable unit; a cluster, an aggregation, or a band" or to be a "number of persons classified together because of common characteristics or community of interest." The difficulty lies in the fact that the dynamic character of the group is lost in the dictionary's definition. Psychologists tend to think of the group as made up of members, each of whom is in close personal relation with every other member. The group implies interaction and communication among two or more individuals whether with, or without, a designated leader but with a tradition of having learned in time and with mutual experiencing to work together. Obviously, such a co-operative amalgamation of individuals who have progressed upward through the stages of organizing, of assaying each other's resources and personalities and of establishing channels of communication into the more significant and higher stages of mutual reinforcement of one another's contributions to solving a problem is not easy to find. Except as described in terms of process observations, the traditioned group has not yet been studied. It is reasonable to assert that such groups have not been utilized in research in groups, either at the elementary or high school levels. And, with the exception of Darley's work with sororities, such groups have not been the basis of research with college students.

The psychologist, however, tends to generalize anticipated values of traditional groups to very many enterprises. Indeed, the psychologist and the educator have failed to recognize that in their research activities, they must necessarily work with groups that are not traditional groups. All of you are familiar with the so-called *ad hoc* group. The *ad hoc* group is usually established by some outside agency (viz., the college teacher as experimenter) to work with, or solve some externally specified task, under the instruction that they do so co-operatively. Such an *ad hoc* group is at least one step away from the more dynamic group. The *ad hoc* group is an artificial assemblage made for the duration of an experiment. Such a group rarely is expected to, and rarely does, continue to develop groupness.

Under an Air Force contract, our research staff has attempted to test the hypothesis that problem solving by *ad hoc* groups or staffs is superior to problem solving of the random individual. Two widely different kinds of problems have been used with two different populations. One type of problem, designated as the field type, is like the familiar O.S.S. task of requiring a cadre of men to cross a mined road without attracting the attention of the enemy and without traces of their route. A series of such problems was given to *ad hoc* teams of college undergraduates to do. In general, the solutions by teams were markedly superior to those of random individuals. Such a result would have been anticipated from the published results of Shaw's well-known experiment in problem solving.

The other type of problem required the solution of a complex human relations difficulty such as taking steps to improve the morale at an Arctic Weather Station under the constraint that the mission of weather reporting must be accomplished. This type of problem was used with officers in the Air Force in the ranks of captain through colonel. The *individual* solutions tended to be as good or better than the average of *ad hoc* staff solutions.

Is it possible that college students work more co-operatively than officers; or is it likely that staffs or teams are superior in those problems for which there is a knowable solution; or is it possible that the responsibility of command makes each member equivalent to a staff? The answers are, of course, unknown. Certainly children in schools today are expected to participate in, and with, groups.

Perhaps the future holds bright promise for group decisions and group problem solving if such experience develops a common element of groupness which is transferred positively to all forms of later group experience.

Edward L. Thorndike would certainly have wished to evaluate the attainment of the objectives. At Teachers College, three attacks are being made on the nature of the differences in the solutions to problems by groups and by individuals. The first is to compare *ad hoc* groups with random individuals; the second to compare *traditional* groups with *ad hoc* groups; and the third is to create mathematical models to account for the group superiority. As a matter of fact, Dr. Solomon and I have developed two such models to consider group behavior in the solution of the Eureka or knowable solution problem. The models are expressed in terms of two hypotheses: Model A, that group superiority is a function only of the ability of one or more of its members to solve the problem without taking account of the interpersonal rejection and acceptance of suggestions among its members; Model B hypothesizes that group superiority is a function only of the pooled abilities of its members. Model B implies that each problem is solved in two or more stages.

The use of Models A and B on the Marjorie Shaw data as well as the data for the O.S.S. field type problem suggests the *acceptability* (but not the proof) of the formulation that the group is superior to the individual *only* because any group has a higher probability of having among its members at least one solver, or several solvers who can independently solve the various stages of the problem. If this be verified by further researches with a broad range of groups and problems, it reopens the question of the "groupness of the group" and the enduring result as a group activity over fairly long spans of time.

Our own experimental evidence contrasting problem solutions by *ad hoc* teams made up of individuals, who had solved the *same* problem previously, supports the generalization that group superiority may be a reflection of the ability of the best one or two members of the *ad hoc* group. In another experiment, an attempt was made to contrast the problem solutions made by "traditioned" and by "*ad hoc*" teams. The traditioned teams comprised members who worked together at the Air University on all group activities as a

staff for three months. The results lead to the inference that the "three-month traditioned group" is equivalent, but not better than, the *ad hoc* group. It must be emphasized that experimental, "traditioned" groups originally had been designated by the authoritarian procedure of selecting individuals from the class at random to make up the group. Hence the "traditioned" group may not have had in it members who would have developed the group cohesiveness that is so essential to a dynamic group with groupness.

No interest group, as sophisticated as educational psychologists, would be tempted to overgeneralize these results. First, the subjects in our experiments were adults who may have overlearned those ways of working out problems that stressed the individual; second, the solutions did not have to be carried into action. While the individual adult, student or officer, may have made a decision or solved a problem as well as or better than the group, this does not suggest that the individual solution must, necessarily, be carried out as well as a group solution or decision; and third, no attempt was made to ascertain what each individual as individual got from the group process of deliberation and closure.

The three limitations to generalizations about individual over group superiority should suggest that educational psychologists must develop hypotheses and the means for evaluating them; the attainment of curricular objectives not only for relatively immediate consequents but also for long time transfer to the tasks and problems that living affords. Indeed, the groupness of the educational psychologist lies not so much in the affiliations with other interest groups; but in the common purposes and needs to get at the facts, generalizations and applications that will lead to better teaching and better learning among themselves and for those that must learn. Groupness for educational psychology is a function of our communications and transactions.

DOROTHY IRENE MARQUART
University of Arizona

Group Problem Solving*

The conflicting adages, "Too many cooks spoil the broth" and "Two heads are better than one," may well serve to illustrate the confusion about a current problem in education: Which is better, individual learning or group learning? In answering this question we have also to consider the free-floating anxiety that we may lack individual resources and are abjectly dependent on the team, or group, and exquisitely attuned to its moods and demands. Such phrases as "organization man" are rather high abstractions with emotional auras that obscure their meaning and maximize their threat. It is never quite clear in what way an "organization man" is different from one who is "non-organizational," especially in a society where only a few artists and country physicians manage a livelihood without much institutional involvement. Even some of our most prominent scientists have become organization men: they work in teams in what must be one of the biggest organizations in the world—the Federal government. It would be a bit of Sophoclean irony if we could produce evidence to show that no child is more organizational than the offspring of child-centered parents eager only to foster his independence. The question may not be conformity versus nonconformity but what type of conformity and who decides the type.

Any investigation of social learning must not ignore the aspects of the cultural milieu reported above. Such research is further complicated by the fact that individual differences rapidly multiply when one is attempting to find both the stimulus and the response variables in the group. Marquart's study has shed some light on this issue. The student may consider the

* Reprinted and abridged with the permission of the author and The Journal Press from the article of the same title, *Journal of Social Psychology*, 41 (1955), 103-113.

following questions: (1) How does Marquart's treatment of the data differ from Shaw's (as described in the report)? (2) How did this alter the conclusions Marquart reached? (3) What might have occurred if she had used more complex problems? Why? (4) What cultural factors may have reduced the quality of the group performance as compared to the individuals'? (5) As revealed in this study, what seems to be the chief characteristic of this situation which accounts for superiority of either group or individual?

A. The Problem

A number of studies have been conducted to determine the effects of various types of social stimulation upon the performance of individuals. We are fairly certain that simple tasks are performed more rapidly when spectators are present than when the subject is alone and that the presence of a co-working but non-competitive group increases the quantity of the work done but decreases the quality of the work. However, few attempts have been made to compare the efficiency of a group and of individual problem solving.

The experiment upon which the present research is based was published by M. E. Shaw in 1932. Miss Shaw compared, as shall we, the performance of individuals and of groups on problems which require more complex thought processes than do those used by Watson. She employed six problems, one-half of which were to be solved by individuals working alone and one-half by individuals working in groups of four members each. She analyzed her data in terms of the total number of correct solutions attained relative to the number of possible solutions and found a very strong advantage for the group over the individual. Her individuals attained correct solutions on 7.9 per cent of their attempts and her groups on 40 per cent of their attempts.

It is the belief of the writer that the method of experimentation employed in Shaw's study is good, that the problems are well selected, but that the method of treatment of the data could be improved upon. Miss Shaw does not consider that group performance is superior to individual performance only if the performance of the

group is better than the performance of the top member of the group. Thus, if one of the members of a group can solve a problem when working alone, there is no proof of the advantage of group performance if the group as a whole can solve the same problem.

The purpose of the present study is to determine whether the advantage of group performance in problem solving will meet this more rigid criterion. Does a group solve problems of the type used by Miss Shaw more readily than does the top member of the group? We shall also reanalyze Miss Shaw's data to check for the attainment of this criterion.

B. Subjects and Apparatus

1. SUBJECTS

The 66 subjects employed in this study were students in a course in experimental psychology at the University of Arizona. They were all of junior, senior, or graduate standing, and most were majors in psychology. All of the subjects had formerly completed at least one year of general psychology and two other courses in the field; most had completed far more than this.

2. APPARATUS

Eight problems were utilized: four for the first experimental session and four for the second. Each problem was typed and enclosed in a separate envelope. The first four problems were either the same or closely parallel to four of those used by Shaw in her study. They were:

1. Use the disks H1, H2, H3, W1, W2, and W3. (For the present disregard the symbols on the reverse side of the disks.)

On the *A*-side of the river are three wives (W1, W2, W3) and their husbands (H1, H2, H3). All of the men but none of the women can row. Get them across to the *B*-side of the river by means of a boat carrying only three at one time. No man will allow his wife to be in the presence of another man unless he is also there.

2. Use the disks marked M1, M2, M3, C1, C2, RC. (The reverse side of the disks just used.)

Three missionaries (M1, M2, M3) and three cannibals (C1, C2, RC) are on the *A*-side of a river. Get them across to the *B*-side by

means of a boat which holds only two at one time. All of the missionaries and one cannibal (RC) can row. Never under any circumstances or at any time may the missionaries be outnumbered by the cannibals. (Except, of course, when there are no missionaries present.)

3. The enclosed slips of paper containing words will when put in the proper order form the last three lines of the unfinished sonnet below. Arrange them as nearly as possible in the proper order.

Knowing this man, who calls himself comrade
mean, underhanded, lacking all attributes
real men desire, that replenish all worlds
men strive for; knowing that charlatan, foot too,
masquerading always in our colors, must also
be addressed as comrade—knowing these
and others to be false, deficient in knowledge
and love for fellow men that motivates our kind.
Nevertheless I answer the salutation proudly,
equally sure that no one can defile it,
feeling deeper than the word the love it bears,
the world it builds. And no man, lying,
talking behind back, betraying trustful friend,
is worth enough to soil this word or mar this world.

4. A consolidated school is to be built in the rural district shown in the diagram. The capital letters (*A, B, C*, etc.) indicate points (not towns) where pupils are to be picked up by two school buses. The mileage between each point is indicated on the diagram. The capacity of each bus is 35 pupils and the driver. Find the most desirable location for this school and give the route each bus must take. The buses may start at *any* point and need not necessarily start from the school each morning. Following are the number of pupils to be picked up at each point (see Figure 1):

Point	Pupils	Point	Pupils	Point	Pupils
A	6	D	4	G	3
B	13	E	2	H	10
C	17	F	5	I	3

The second set of problems were as follows:

1. A man stopping at a hotel ran out of money and would have no more for 23 days. The landlord would not trust him, but, as he

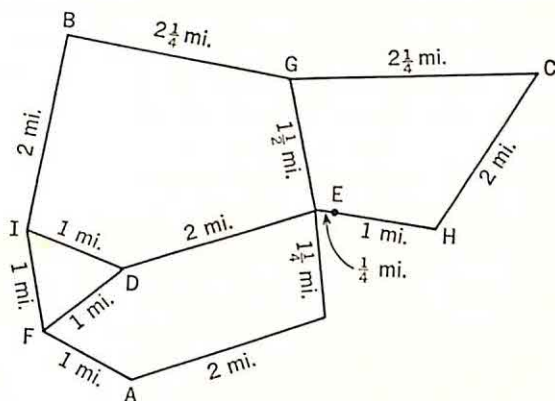


Fig. 1

had a heavy gold chain of 23 links, the landlord agreed to accept one link in payment on each successive day and to restore the chain on receipt of the money. The owner was anxious to preserve the chain as intact as possible. How many links was it necessary for him to cut? Show how he was able to pay the landlord one link on each successive day for 23 days, and yet cut only this small number of links of the chain. (Time limit 30 minutes.)

2. Six bathing beauties stood in line facing the judges.

The prize was given to the only girl whose name began with the same letter as that of her state; only, of course, the judges did not know this until the contest was over.

From the description below, you will be able to write each girl's name opposite her state, and then pick out the prize winner.

Miss Ohio was not on speaking terms with *Dorothy*

Olga was engaged to *Miss Delaware's* Brother.

Mary and *Miss Maryland* were at opposite ends.

Dorothy was at the judges' right, next to *Miss Maine*.

Neither *Maude* nor *Vera* represented *Ohio*.

Miss Vermont was between *Katie* and *Miss Delaware*.

Miss Kansas was between *Olga* and *Miss Maine*.

Vera was not next to the girl at the judges' left.

(Time limit 25 min.)

3. A milkman has a 14-quart can full of milk. He wishes to divide his milk into two equal portions. In addition to the 14-quart measure, he has a 5-quart measure and a 9-quart measure. How does he make the division without any waste using the three measures only and not guessing at the amounts? (Time limit 30 min.)

4. In the first four tricks of a four-handed game of cards, there was only one King, which was the highest card played. There was only one Six, which was the lowest card played.

No two cards of equal value were played in any single trick.

Find the value of all four cards in each of the four tricks. The following statements will give you the information you need.

a. The highest card in the 1st trick was equal in value to the 2nd highest card in the 4th trick.

b. The lowest card in the 2nd trick was equal in value to the highest card in the 4th trick.

c. The second highest card in the 2nd trick was equal in value to the lowest card in the 3rd trick.

(Time limit 25 min.)

C. Procedure

The 66 subjects employed in the present study were divided into 22 groups of three subjects each.¹ This division, except for an equation of sex, was done on a chance basis. All groups except three were comprised of two men and one woman. These three groups were made up of two women and one man.

Two experimental sessions were utilized for each subject. The first session, of approximately two and one-half hours, was devoted to the solving of the four problems which were identical or closely analogous to four of those used by Shaw. The subjects solved two of the four problems by the group method and two by the individual method. As can be seen in Table 1, one-half of the subjects worked first as individuals and then as groups; one-half worked first as groups and then as individuals. Also, one-half of the subjects worked on Problems I and II first, and one-half of the subjects worked on Problems III and IV first. No time limits were set for the solution of these problems. When working as individuals, the subjects were

¹ There were only 63 subjects divided into 21 groups employed in the second problem solving period. One member of the 22nd group was absent and, although a subject was present to replace him, the data for this group has been discarded.

together in one large laboratory room; when working as groups, each group used a small room.

The individuals, when working alone, were told:

You will be given two problems; solve them in order. Do not be content with your first solution but work until you are fairly certain that you have the correct answer. Record tentative and final solutions on paper so that they can be turned in and scored. Keep a record to the nearest minute of the time required for the solution of each problem. Include the time required for the writing of the solution. If you find it impossible to solve any problem, you may turn in a sheet containing a description of your partial solution.

The groups were told:

Solve the two problems in order. Do not be content with your first solution but work until you are fairly certain that you have the correct answer. One of you is to act as a recorder, that person is to record tentative and final solutions on paper so that they can be turned in and scored. Also, keep a record of the time required for the solution of each problem. Include the time required for the writing of the solution. The recorder is to take an active part in the solution of the problems. Please try to work together as a group as well as you can. Do not work in a parallel manner and then select the best of the three solutions. If you find it impossible to solve any problem, you may turn in a sheet containing a description of your partial solution.

The experimenter observed brief sections of the working periods of the individuals and of the groups to assure herself that the directions were being followed.

The second experimental session, which was held one week later, was devoted to the solving of the second group of four problems. Again, one-half were solved by the individual method and one-half were solved by the group method. The groups were the same as those used during the first session and followed the same procedural order (see Table 1). The only difference is that Group XXII was broken by an absence, and that data for it was discarded in this part of the study.

All directions for this part of the experiment were given while the subjects were together in groups of approximately 20 members. The subjects were told that they were to follow the same procedure as the previous week, that they were each to solve two problems as individuals and two as groups, and that they were to be even more careful to do their group work as a unit. They were told that the

Table 1

PROCEDURE FOLLOWED BY EACH OF THE GROUPS EMPLOYED
IN THIS STUDY

Group	First one-half of session		Second one-half of session	
	Problems	Method	Problems	Method
I	1 and 2	individual	3 and 4	group
II	1 and 2	group	3 and 4	individual
III	3 and 4	individual	1 and 2	group
IV	3 and 4	group	1 and 2	individual
V	1 and 2	individual	3 and 4	group
VI	3 and 4	individual	1 and 2	group
VII	1 and 2	individual	3 and 4	group
VIII	1 and 2	group	3 and 4	individual
IX	3 and 4	individual	1 and 2	group
X	3 and 4	group	1 and 2	individual
XI	1 and 2	individual	3 and 4	group
XII	1 and 2	group	3 and 4	individual
XIII	3 and 4	individual	1 and 2	group
XIV	3 and 4	group	1 and 2	individual
XV	1 and 2	individual	3 and 4	group
XVI	1 and 2	group	3 and 4	individual
XVII	3 and 4	individual	1 and 2	group
XVIII	3 and 4	group	1 and 2	individual
XIX	1 and 2	group	3 and 4	individual
XX	1 and 2	individual	3 and 4	group
XXI	3 and 4	group	1 and 2	individual
XXII	3 and 4	group	1 and 2	individual

only difference in procedure from that of the previous week would be that there would be a time limit of 30 minutes for Problems I and III and of 25 minutes for Problems II and IV, that they were to use no more than this time per problem, and that they were to examine their solutions carefully before they decided that the solutions were correct and ready to be turned in.

D. Results of the First Experimental Session

When the results of this study are analyzed in the same manner as that used by Shaw, her findings are corroborated. That is, if one considers the ratio of the number of solutions obtained to the number of possible solutions, the groups solved many more problems than did the individuals. Since there were 66 subjects and each worked on two problems individually, 132 solutions were possible.

Twenty-nine solutions were obtained if Shaw's criterion for the answer of the fourth problem is used. The author questions whether an answer which makes the two buses travel a total of one-half mile farther is not as good or better than her solution.² If the solutions allowing the buses to travel this additional one-half mile are allowed, the ratio of solutions to attempts is 35 to 132.

Forty-four problems were attempted by groups. Using Shaw's criterion, there were 17 successes; using our more lenient requirement (either 12 or 12½ miles for Problem IV), there were 19 successes. The ratio of successes to attempts is either 17 to 44 or 19 to 44.

There seems to the writer to be a question of the validity of this statistical treatment. It counts one individual as equal to one group and makes no allowance for the fact that a group solution might be the result of activity of any one of the three individuals rather than of the group as a whole. One can reanalyze the data for individual solutions by viewing the individuals as grouped. If any one, or more than one, individual in the group solves the problem, the group working individually can be given credit for the solution. For example, when working as individuals, one member of Group I solved Problem I. According to this method of calculation, Group I working as individuals solved Problem I. On the same basis, two members of Group IX solved Problem I and the group is given credit for an individual solution of Problem I. No member of Group IV solved the problem; thus the group working as individuals is credited with a failure on the problem. When this is done, 61 per cent of all problems attempted by the groups working as individuals were correctly solved.³ Only 43 per cent of all the problems attempted by the groups, as groups, were solved.

This difference between solutions by 61 per cent of the groups working individually and by 43 per cent of the groups working as

² Shaw's solution probably places the school between *B* and *G*, 1¼ miles from *G*. By this method the two buses travel absolutely equal distances. However, one of the buses will have to start its route by picking up the 17 children from *C*, thus causing the pupil distance travelled to be large. By adding the additional one-half mile to the distance travelled and placing the school at the same point the large ensembles of children can be picked up last rather than first.

³ This value is based upon the acceptance of a distance of either 12 or 12½ miles as correct for the fourth problem.

groups is not significant. The chi-square value obtained from using the test of independence is 3.684 with one degree of freedom. This value attains the 5.7 per cent level of significance.

When one considers the problems separately, there are some slight differences in trends (see Table 2). Problems I, III, and IV were solved slightly more readily by individuals than by groups. Problem II was solved slightly more readily by groups than by individuals. Only one of these differences even approaches significance. The difference for Problem III between the proportions of groups solving as individuals and of groups solving as groups attains the 3.5 per cent level of significance when the chi-square test of independence is employed.

The average time required for correct solutions by the groups working as groups is slightly higher than the average time required for the groups working as individuals⁴ (see Table 2).

Table 2
RESULTS OF THE FIRST EXPERIMENTAL SESSION VIEWED WITH
INDIVIDUALS WHEN WORKING ALONE STILL CONSIDERED
AS GROUPED

Problem	Per cent solving correctly		Average time required for solutions*	
	Individuals	Groups	Individuals	Groups
I	75%	50%		
II	25%	50%	9.2 min.	6.6 min.
III	80%	33%	13.3 min.	31.0 min.
IV (12 miles)	10%	17%	8.3 min.	15.8 min.
(12½ miles)	50%	17%	14.0 min.	19.5 min.
Total	61%	43%	11.0 min.	15.5 min.
			10.8 min.	15.5 min.

* The average time required for solution is an average of the times required for each group to solve the problem correctly. Incorrect solutions are disregarded.

E. Results of the Second Experimental Session

The results of the second experimental session are very similar to those of the first session. Using Shaw's method of treatment of the

⁴ The average time required for correct solution by groups working as individuals was attained by averaging together the lengths of time required for a correct solution by the most rapid worker in each group.

data and allowing as correct a cutting of four or fewer links in Problem I (2 only need be cut), 36 out of 126 possible correct solutions were attained by the individuals, 22 out of 42 possible correct solutions were attained by the groups.

When one analyzes the data in terms of groups working individually as compared with groups working as groups, there is a slight advantage for the individual rather than the group method of solving problems. Sixty per cent of the solutions attempted by groups working as individuals were attained, while 52 per cent of the solutions attempted by groups working as groups were correct. This difference is not significant.⁵

Again there is a variation between the problems (see Table 3). Problems I, III, and IV were solved slightly more often by groups working as individuals than by groups working as groups. Problem II was solved slightly more frequently by groups than by individuals. None of these differences even approach significance.

The average time required for a correct solution was, in this case, slightly greater for the groups working as individuals than for the

Table 3
RESULTS OF THE SECOND EXPERIMENTAL SESSION VIEWED WITH
INDIVIDUALS WHEN WORKING ALONE STILL CONSIDERED
AS GROUPED

Problem	Per cent solving correctly		Average time required for solutions*	
	Individuals	Groups	Individuals	Groups
I (2 links cut)	0%	10%	—	10 min.
(3 links cut)	9%	10%	30.0 min.	20 min.
(4 links cut)	36%	10%	25.2 min.	6 min.
II	27%	40%	22.3 min.	24 min.
III	80%	73%	17.7 min.	22.4 min.
IV	90%	64%	13.4 min.	17.4 min.
Total	60%	62%	20.25 min.	18.95 min.

* The average time required for solution is an average of the times required for each group to solve the problem correctly. Incorrect solutions are disregarded.

⁵ If one demands absolute perfection in answer before crediting a group with success (that is, if one allows only two links to be cut in Problem I), 48 per cent of the solutions attempted by both the groups working as groups and by the individuals viewed as groups were attained.

groups working as groups. However, as can be seen in Table 3, this trend is due entirely to differences in the time required to solve Problem I.

F. Summary and Conclusions

We have believed for some time that group performance is better than individual performance, that one cannot simply add together the results of the production of the individual members of a group and arrive at a group production. We claim to have demonstrated this phenomenon in our psychological laboratories. And we group our scientists together in order to assure the accomplishment of the scientific aims of our period.

The writer feels that the results of the experiments which have been performed to demonstrate group superiority are inconclusive. This does not mean that our national research accomplishments do not seem to reveal an advantage of group performance over individual performance. Watson found that a group can construct more words out of the letters found in a longer word than can the best member of the group. He also found that the group performance is positively related to the individual performance of the best member of the group ($r = .53$), negatively related to the performance of the poorest member of the group ($r = -.19$), and only negligibly related to the performance of the average number of the group ($r = .16$). Does this not sound as though the results are due to a simple summation? Considering this and the nature of the task, it would seem that for the group to be regarded as superior to the members of the group working individually, the performance of the group would have to surpass the total production of the individuals. This was accomplished by only five out of 20 groups.⁶

Shaw found that in the solving of problems requiring reasoning the average group surpassed the average individual. She did not check to determine whether the performance of the group is superior to the performance of the top member of the group. If one makes such an analysis from her data, one finds that 12 of the 30

⁶ Watson also gave his groups an advantage by adding a secretary (not a member of the group) to record the responses of the group. The individuals when working alone recorded their own responses.

solutions attempted by the groups were correct and that 9 of the 30 solutions attempted by the groups working individually were correct. This gives an advantage, although very slight and not significant, to the group over the top member of the group.⁷

Our results are similar to those obtained by Shaw. If we analyze our data in the same manner as does Shaw, we find an advantage for the group over the individual. But when we analyze our data to determine whether the group results surpass the individual results of every member of the group, we find no significant difference. The trend, in fact, favors the top individual of the group over the group as a whole.

Perhaps the advantage of the group over the individual members of the group appears only when the problems are even more complex. That is, when the amount of background work that must be done is so great that no one individual can possibly accomplish it. If there is a difference in the problem solving behavior of groups and of individuals using the type of problem employed by Shaw and in the present study, it must be in approach rather than in the merits of the solution. The only alternative conclusion would be that our pleas for co-operation were not followed and that superiority can appear only if our training methods are changed in such a way as to increase the co-operative nature of our culture and to decrease its competitive nature.

References

- SHAW, M. E., A comparison of individuals and small groups in the rational solution of complex problems. *Amer. J. Psychol.*, 1932, 44, 491-504.
- WATSON, G. B., Do groups think more efficiently than individuals? *J. Abn. & Soc. Psychol.*, 1928-29, 23, 328-336.

⁷ If one makes the criterion of the school bus problem more lenient and allows the travelling of the additional one-half mile, the two values are identical. Both the groups working as groups and the groups working as individuals solved 13 problems.

W. J. McKEACHIE
University of Michigan

Student-Centered Versus Instructor-Centered Instruction*

Lorge referred to the "Charybdis of permissiveness and the Scylla of control" as they relate to the use of group instruction in the schools (p. 532). "How unstructured (that is, permissive) is Mr. Arthur, the biology teacher?" is the sort of question that is asked in school and college halls. If it is discovered that Mr. Arthur is quite "unstructured," his colleagues (but not necessarily the principal) are relieved and feel they can entrust students to him. If they find that he is in full control not only of his materials and his emotions but also of the class, they may begin investigating, not the situation, but Mr. Arthur's early childhood experiences. This is hardly research.

Partly to remove the issue from the realm of stereotyped thinking, McKeachie has attempted to define some of the research problems connected with student-centered versus instructor-centered teaching. The following questions are raised to help the student compare McKeachie's ideas with previous readings and comments: (1) How can student participation in goal setting be reconciled with the school's responsibility to social demands (see Introduction, p. 528)? (2) What advantages in motivation may group-centered instruction have over the other method? How would you test your hypotheses to find out? (3) Why would the instructor-centered method have advantages for reinforcement which may be lacking in the other? (4) How can the type of independent discovery de-

* Reprinted with the permission of the author and the Abraham Magazine Service, from the article of the same title, *Journal of Educational Psychology*, 45 (1954), 143-150. Footnote and references are omitted.

scribed by Bruner, Kersh, and others occur in the instructor-centered class? (5) What might explain the students misperceiving how well they have done in achieving either cognition or attitudinal goals?

Even psychologists have their stereotypes. And for most of us "student-centered" and "instructor-centered" are stereotypes. With "student-centered" we associate the halo terms of democratic, permissive, insight, affective and student growth. "Instructor-centered" brings to mind the terms authoritarian, Fascistic, knowledge for its own sake, and content-centered. In our psychological subculture the mere labels in our title stack the deck against anyone who attempts to defend the instructor-centered point of view.

Despite our preconceptions, there are important areas of agreement; e.g., everyone agrees that we're teaching students and that our job is to promote student learning. Also, almost everyone agrees that we want to improve our students' problem-solving skills. This paper aims to do three things:

- 1) To break up our stereotypes by examining some of the dimensions which may differentiate student-centered from instructor-centered instruction.
- 2) To survey a group of research studies on the problem.
- 3) To do some theorizing about the problem.

Dimensions of Difference

First, what are the dimensions of difference?

Goals

One of the most prominent is the dimension of goal setting. The instructor-centered teacher believes that he is ultimately responsible for determining goals. To quote the report of a study group which met at Cornell last year, "If the teacher merits the responsibility placed in his hands, he knows more than do the students about the subject, about the world in which we live, and of the ways in which

a knowledge of psychology can enrich the world." The student-centered instructor, on the other hand, believes that the group, including both students and instructor, should determine the group goals.

Despite his greater emphasis upon the student in goal-setting, the student-centered instructor usually has certain implicit goals which he hopes will be achieved by the students. Thus another area of difference between instructor-centered and student-centered teaching is in the type of goals for which they are aiming. This area of goals needs to be plotted in at least two dimensional space because there are three differing types of goals: intellectual, applied, and affective.

One approach emphasizes the traditional intellectual goals of a liberal education. It attempts to create an interest in "Knowledge for its own sake." The primary goal is to teach students to think. The instructor with this approach may be interested in attitudes, but they are the attitudes of the scientist toward his subject matter, not social attitudes. These are the goals which we usually associate with the instructor-centered method, but which may be independent of other dimensions of the methods.

The instructor who is interested in applied goals feels that psychology can offer much that is useful in the student's daily life, not only in adult life but in college as well. He is interested in teaching the facts of psychology only insofar as they can be applied by the student. Specific skills such as reading skills or study habits are apt to be part of the content of his course.

The instructor primarily interested in affective goals is likely to disavow any suggestion that his class is simply group therapy. In spite of this, he is even more dissatisfied than most teachers with achievement tests as a criterion of his teaching, and he is quite unhappy if his students don't have some adjustment problems that their psychology course can cure.

Student-centered instruction tends to emphasize the last of these three goal-emphases, but none of these goal-emphases are insolubly linked with other dimensions of teaching. Benjamin Bloom, for example, has been much interested in the discussion method as a technique for attaining intellectual goals.

Methods of Teaching

Probably most of us think of student-centered instruction as differing from instructor-centered teaching primarily in terms of what goes on in the classroom. One of the dimensions here is degree of student participation. The student-centered instructor tends to encourage a high degree of student verbal participation. Most student-centered instructors, however, are not satisfied if student discussion is directed at the instructor. They are interested in developing a high degree of inter-student participation. The student-centered instructor feels that students who talk to him are maintaining their dependence upon him, and this is in conflict with his usual goal of independence.

Another dimension of classroom behavior is the degree of instructor acceptance of erroneous or irrelevant student contributions. Some student-centered instructors emphasize the importance of accepting all contributions without evaluation, or at least without negative evaluation.

A third dimension of classroom climate is degree of group cohesiveness. Typically the student-centered instructor attempts to create a group with a high degree of cohesiveness. He may attempt to measure this by counting the number of "We's" as contrasted with the number of "I's" verbalized in class discussion.

A further dimension upon which student-centered and instructor-centered classes differ is in the degree to which the student feels he can influence his own fate. Perhaps it is unnecessary to elaborate upon this point, but it may be pointed out that what have been called student-centered classes vary from classes in which the instructor lays out a course outline, makes assignments, and actively guides discussion, to classes in which almost all course planning is done by group decisions and the instructor begins class with, "What would you like to talk about today?"

The amount of class time devoted to personal experiences and problems of the students may be another dimension. The instructor who is concerned with the development of self-insight is more apt to permit or encourage discussion of personal problems.

Obviously one could list many other dimensions upon which instructor-centered and student-centered teaching may differ. Those

listed are simply some which have been most frequently mentioned in the literature. When someone says he has used student-centered teaching he usually means that as compared with instructor-centered teaching, his class has had a higher degree of one or more of these qualities:

- 1) Student participation in goal setting.
- 2) Emphasis upon affective goals.
- 3) Student participation and student interaction.
- 4) Instructor acceptance of inaccurate statements.
- 5) Group cohesiveness.
- 6) Ability to determine its own fate.
- 7) Amount of time devoted to discussing personal experiences and problems.

Experimental Results

One would expect that the controversy between our education's authoritarians and permissivists would long ago have been resolved by the cold logic of experimental studies. Unfortunately, this just hasn't happened. The published experimental studies are not in agreement and there are a host of unpublished studies which remain unpublished because the two methods used produced no significant differences in outcomes.

By way of some examples: One of the best known comparisons of student-centered and instructor-centered instruction is that of Faw. Faw's class of one hundred and two students met two hours a week to listen to lectures and two hours a week in discussion group of thirty-four. One of the discussion groups was taught by a student-centered method, one by an instructor-centered method, and one group alternated between the two methods.

As compared with the instructor-centered class the student-centered class was characterized by more student participation, no instructor correction of inaccurate statements, lack of direction, and more discussion of ideas related to personal experiences.

Surprisingly enough, Faw's major measure of attainment of objectives was in the intellectual area. Scores on course examinations showed small but significant differences favoring the student-cen-

tered method. In the area of his major interests—emotional growth—Faw's method of evaluation was to ask students in the student-centered and alternating classes to write anonymous comments about the class. Generally these comments seemed to indicate that the students felt that they received greater social and emotional value from the student-centered discussion groups than they would have from an instructor-centered group. Despite the objective test results Faw's students felt that they would have made greater intellectual gains in an instructor-centered class.

Now compare Faw's experiment with that of Asch. Asch, like Faw, taught all of the groups involved in his experiment. Three sections of about thirty to thirty-five students were taught by instructor-centered methods; one section of twenty-three students was taught by a student-centered method, quite similar to that of Faw. However, there were certain differences between Faw's and Asch's experiments. In Faw's experiment both student-centered and instructor-centered classes spent two hours a week listening to lectures. In Asch's experiment, only the instructor-centered classes were subjected to lectures. While Faw doesn't mention grading, one assumes that grades were determined by the instructor on the basis of the course-wide examination. In Asch's experiment students in the student-centered class were allowed to determine their own grades.

The interesting thing is that Asch's results do not completely agree with Faw's. On the final examination in the course students in the instructor-centered classes scored significantly higher than members of the student-centered class, not only on the objective portion of the test but also on an essay portion. Note, however, that the student-centered class was specifically told that the examination would in no way affect their grades in the course so that these differences may be simply due to a difference in motivation. As measured by the Bogardus Social Distance scale, attitude change in the two sections was not significantly different. However, as compared with the instructor-centered classes a greater percentage of members of the student-centered class improved in adjustment as measured by the MMPI.

Interestingly enough Asch's students, like Faw's, had a different perception of their achievement than that shown by the course examination. Faw's student-centered class did better on the course ex-

amination than the instructor-centered section, but thought they would have learned more if they had been in an instructor-centered class. Asch's students rated the student-centered class higher than instructor-centered in helping them to learn the subject matter of the course but they actually scored lower. There seems to be some irony in the fact that advocates of student-centered methods find that students' perceptions of group achievement is erroneous; yet many of us want students to take a larger share of the responsibility for evaluation, and are pleased that they report great gains in personal and social values in student-centered classes. If groups which report greater intellectual gains actually learn less, it might seem logical to conclude that groups which report greater gains in the areas of personal emotional growth may really be gaining less than those groups which report lesser gains.

One of the most comprehensive experiments in this area is that of Landsman. He experimented with a student-centered type of teaching as contrasted with a more directed type of democratic discussion organized around a syllabus. His experimental design involved eight classes in a course sequence of Human Development, Adjustment, and Learning. Three instructors took part in the experiment and each instructor used both methods. Outcome measures included the Horrock-Troyer tests, a local case history analysis test, Group Rorschach, MMPI, autobiographies, and students' reactions. His results showed no significant differences between methods.

A Redefinition of the Problem

What are we to conclude from these studies? While there is a dearth of follow-up data, with such slender results at the end of the courses our hope that either method produces significantly greater long-time benefits is probably unrealistic. Should everyone go his own way and teach any way he pleases? Personally, we are not willing to go quite so far, but certainly none of us should exclaim with horror, "His classes are instructor-centered!"

As psychologists, however, we believe in research. Why has research on student-centered versus instructor-centered teaching seemed to lead up a blind alley?

One reason suggested for contradictory results is that different people have meant different things by student-centered. But a far

more important reason is that we've been lumping together more variables than we could handle with our experimental designs. The first part of this paper was concerned about defining dimensions of difference because we need to work with a more limited number of variables and we need to relate these variables to the main body of psychological theory.

Do we have any assurance that in the welter of student-centered versus instructor-centered research there are any variables that make a difference? Despite a somewhat pessimistic view of most research in this area, there are enough glimmerings of hope to justify further research.

For example, Smith and Johnson have found that student-centered teaching produces higher scores than instructor-centered teaching on tests of reasoning ability and creativity. Furthermore, all research on the problem seems to agree that as compared with instructor-centered teaching, student-centered teaching results in little decrement to the learning of facts (providing the classes have textbooks and tests are based on the texts).

In addition Asch and Faw are not alone in feeling that adjustment and social skills are improved by student-centered methods. The research of Gibb and Gibb indicates that students from group-centered classes which possessed many of the characteristics ordinarily called "student-centered" actually produced a growth in social skills in experimental situations outside the classroom. Kelley and Pepitone found an increase of empathy in classes which had been taught by student-centered methods. Efforts to produce such gains in instructor-centered classes have been unsuccessful.

Obviously, we cannot test every possible interrelationship of the variables involved in student-centered classes, but many can be studied independently of the total method. For example, the variable of degree of student verbal participation in the classroom could be divorced from goals, cohesion, and other variables, and conceptualized as providing an opportunity of student responses to be rewarded or punished. We might gain some insight into student-centered teaching by varying opportunities for verbal participation or by varying the percentage of comments which the instructor rewards or corrects.

Another variable upon which we already have data is the degree to which the student feels able to influence his own fate. It seems

significant that almost every one who tries student-centered teaching methods finds that the problem of grades presents the greatest obstacle to success. In addition two studies which produce evidence of change as a result of student-centered teaching both reduced the power of the instructor by giving the group responsibility for determining grades.

Our own research at the University of Michigan supports our notion that the student's feeling of freedom is an important variable. In an earlier paper the writer reported that students made higher scores on examinations when they were given the opportunity to write comments about test items. More recent experimentation indicates that the effect of this opportunity is reduced if the student is directed to write comments. Apparently the important thing is that the student feel free to do so if he wishes to.

In addition, it was found in another study that students preferred a directive method of teaching which made clear what the student had to do in order to pass the course. Again, we would interpret this as meaning that the student felt better able to determine his fate.

Looked at in this way, the class in which the instructor insists on being non-directive may actually increase the student's feeling of helplessness since he doesn't know what to do in order to achieve the goal of a good grade in the course. In this situation the student may simply perceive the instructor as using his power to block the normal pathways to the goal.

From this theory one would predict that the effect of instructor permissiveness would depend upon whether or not the group possessed the skills necessary to achieve their goals. In a new group the effect of instructor permissiveness may depend upon the presence or absence of individuals in the group who have had previous experience in working in democratic groups. If the instructor retains control of rewards, permissiveness with respect to means to the goal (such as assignments, classroom activities, etc.) may simply increase the ambiguity of the situation for the student and reduce student learning.

Obviously such speculations need to be tested. To make our research productive we need to sharpen both our theories and our methodology.

References

- ASCH, M. J., "Nondirective teaching in psychology." *Psychol. Monogr.*, 1951, 65, 4.
- FAW, V., "A psychotherapeutic method of teaching psychology." *Amer. Psychologist*, 1949, 4, 104-109.
- GIBB, LORRAINE, and GIBB, J., *The Effect of the Use of "Participative Action" Groups in a Course in General Psychology*. Paper at 1952 APA meetings.
- KELLEY, H., and PEPITONE, A., "An evaluation of a college course in human relations." *J. educ. Psychol.*, 1952, 43, 193-209.
- LANDSMAN, T., *An Experimental Study of a Student-centered Learning Method*. Unpublished doctor's dissertation, Syracuse Univer., 1950.
- MCKEACHIE, W. J., "Anxiety in the college classroom." *J. educ. Res.*, 45, 1951, 153-160.
- SMITH, H. C., and JOHNSON, D. M., *An Experimental Study of Attitudes and Achievement in the Democratic Classroom*. Paper read at 1952 APA meetings.
- WOLFLE, D., et al., *Improving Undergraduate Instruction in Psychology*. New York. Macmillan, 1952.

ALVIN ZANDER, *University of Michigan*

ARTHUR R. COHEN, *Yale University*

Attributed Social Power and Group Acceptance: A Classroom Experimental Demonstration*

Some aspects of social interaction are not always pleasant to know about. In the "enlightened" twentieth century we would like to believe that, with the exception of international politics

* Reprinted with the permission of the senior author and the American Psychological Association, from the article of the same title, *Journal of Abnormal and Social Psychology*, 51 (1955), 490-492.

and certain lapses on the domestic scene, the grim realism of power politics is as much a relic of the past as the madrigal. Psychologists, who are usually polite middle-class individuals, have been somewhat averse to studying power. They have been less reluctant to study conformity, a morally acceptable but not always praiseworthy aspect of middle-class life, than to study the amoral use of power and the powerful nonconformists. The usual research faintly suggests that the independents or nonconformists are detached, self-orbiting entities. But this hardly explains a Napoleon, a Caesar, or even a college dean.

The following demonstration illustrates that power stalks over the social stage. The study is an example of the *ad hoc* groups described by Lorge (pp. 534-536). It is also a study of the psychological effects of verbal behavior and communication (see Lorge, pp. 326-338). It reveals the workings of a social hierarchy of humans. Since the "powerful person" is a source of reinforcement, while the erstwhile freshman seemingly has no reward value, we see the operation of reinforcement in a group setting.

The student may ask: (1) How might these results have differed if the group were real instead of *ad hoc*? (2) What similarities does this experiment have with those of Verplanck on operant conditioning (pp. 140-156)? (3) What implications does this have for the role of the teacher in the classroom if we substitute a teacher for the dean? (4) In psychological terms, what would a teacher have to do if she wanted the students to be more independent of her attributed power?

This paper describes a classroom demonstration which illustrates the reaction of group members to the communications and behavior they receive from others, when the nature and direction of the communications are influenced by the amount of power attributed to special members.

The exercise is based on the assumption that individuals are likely to be sensitive and alert toward persons to whom they at-

tribute much power, and relatively less concerned with those who are viewed as having little power. Attributed power is here defined as the perception of a person's ability to influence or determine the fate of the perceiver whenever the powerful person wishes to do so. The unique feature of power, as distinguished from prestige, authority, or status (and the probable cause of the sensitivity it generates in subordinates), is that the powerful person is typically perceived as a potential source of either need gratification or deprivation by those with little power. The latter individuals in turn behave toward superiors in a fashion which maximizes the possibility that the powerful persons will facilitate the subordinates' need gratification.

Such alertness has been noted in various studies. The more that power is attributed to a person, the greater is the likelihood that he will be the target of communication from subordinates, receive more deferential or solicitous behavior from them, and be perceived as friendly toward the observer. Assuming that the subordinates are not able or willing to avoid interaction with the one to whom they attribute high power, it has been reported that their behavior toward the superior may be either an attempted substitute for upward mobility, or an effort to win support and rewards from the superior.

Since deferential, solicitous, and attentive behavior is more often directed toward persons with high attributed power, it may be derived, other things being equal, that a member who is treated as though he were a powerful person will have greater attraction to the group and higher self-evaluation than one who is treated as though he were a subordinate. He will be attracted to the group because he views his position as secure, and because he feels wanted and valued by the group. His self-evaluation will be positively influenced by the same perceptions.

The demonstration is a procedure for illustrating these notions. It precedes any class discussion of power, group structure, position, or related concepts and serves as an insight-producing experience which provides a readiness to explore such subject matter. The exercise is so designed that the persons to whom high and low power is attributed are not aware that they have a unique position in the eyes

of the others in the group. This is done in order to prevent the development of any confounding expectations arising from perceptions of own power.

METHOD

The students are instructed to form into groups of seven persons each. They are told in somewhat more detail than is necessary here that each circle is to think of themselves as constituting a committee which has been appointed by the vice-president of the University and to ignore the existence of the other similar committees in the room. Each group has the hypothetical task of preparing advice for the vice-president as to the most suitable policy for the use of a large sum of money which has been given to the school by an anonymous donor. The man who provided the funds has stipulated only that students have at least some voice in determining how the money should be expended, and that the money be used for student welfare. They are informed that each committee is to assume that it has already met once prior to this occasion but at that time felt that their group was too small and unrepresentative of the relevant viewpoints. Thus, they have asked the vice-president to appoint several additional members to their group.

Each subgroup is then requested to select two persons to send out of the room. These are to be the new members just appointed by the vice-president at the committee's request. This step makes it possible to state a reasonable but false purpose by asserting that the objective of the demonstration is to show how it feels to be a newcomer to a group which has been previously organized. The persons sent into the hall are given further instructions, asked to select a partner different from the one with whom they had left the room, and to return on signal to a group different from the one they had both left. They are told that they will be interviewed in front of the class, after their meeting with a committee, concerning how it feels to be a new member. They are not informed that they will have unusual social positions attributed to them when they arrive at the meeting, nor is anything said about the behavior expected of them by their groups. Thus, in effect, they are induced to be students in the role of late arrivers to an organized group.

Meanwhile the groups in the classroom are given the information that the vice-president has selected two new committee members in accordance with their request, that one is the Dean of the School of Education (the subjects are mostly Education majors) and the other is a college freshman. They are instructed to be very sure that they do not tell the newcomers about these positions in the subsequent meeting. They are asked to decide as a group which location in the circle

(when occupied by a new member) will be considered to be the Dean and which the freshman, and are given a few minutes to develop some history so that they can properly act as though they have had a meeting prior to the arrival of the new members. They are not told to treat the newcomers differently, nor is anything said as to how they should act toward the new members.

The persons who have been sent out of the room are then asked to join the committee of their choice and the meetings begin. Typically the "Deans" are received with much interest and attention by the groups, while the "freshmen" are greeted with more detachment and given less notice. The discussion is usually directed toward the newcomers in large part, but primarily to the Dean. After 10 minutes the discussion is stopped by the instructor and all class members are asked to fill out a brief questionnaire which is introduced as a means to gather their private thoughts about the meeting. The completed questionnaires are then set aside for the moment and various newcomers are interviewed before the class concerning their feelings while they had been participating as group members. These oral reports illustrate for the students that the Deans and the freshmen react differently to the group experience. The Deans tend to emphasize the attractive qualities of the group and the interest the members displayed in his ideas. The freshman's remarks are quite the opposite.

The demonstration is concluded by simultaneous discussions within each group. During this time the manipulation is revealed to the newcomers in large part, but primarily to the Dean. After 10 minutes the two levels, and the members discuss the reasons for their behavior toward the Dean and the freshman in their group. Finally, before the questionnaires are handed in, each participant is asked to indicate on it whether he had been a Dean, a freshman, or a group member.

QUESTIONNAIRE RESULTS

Data concerning reactions to the meeting are available from the reflection questionnaires completed by the Deans and the freshmen in 21 subgroups before they had been told of their unwitting roles and prior to any group discussion concerning the nature of their interpersonal relations. This instrument, which is very brief for practical reasons, contains five graphic rating scales (eight points). On two of the questions there are data from only 13 groups since these items were added after the demonstration had been tried in one class containing eight groups.

It is apparent in Table 1 that the Deans felt attracted to their group, perceived that they had made a good first impression, that

Table 1
AVERAGE REACTIONS OF "DEANS" AND "FRESHMEN" TO
THEIR GROUP EXPERIENCES

Measure	N in each role	Deans Mean	Freshmen Mean	<i>t</i>	<i>p</i>
Attraction of group for member	13	6.2	4.1	3.79	.001
First impression made on group	13	5.0	3.9	2.41	.03
Social validity given to opinions	21	5.4	3.9	2.63	.02
Degree of influence on group	21	6.8	4.4	4.49	.001
Ease felt in group	21	5.8	5.2	.87	.20

the group agreed with their ideas, and believed that they had had a stronger influence in the group than did the freshmen. All of these differences are statistically significant at acceptable levels of confidence. The two new members were not significantly different in the degree to which they were at ease during the meeting, although the findings are in the direction one might expect.

These results may be interpreted to mean that the Deans were more attracted to the group and evaluated themselves more highly than did the freshmen. They suggest that the group members tended to treat the two newcomers in quite different fashions since this is the most probable cause of the discrepancy in the reactions of the late arrivers.

Experience with this demonstration (and others like it) in classrooms and audience settings indicates that persons will readily differentiate in their behavior toward others to whom they are asked only to attribute some characteristic such as high or low power. In regard to the person whom they perceive in a high power role they will be attentive and appreciative, or appear to feel that way. They will appear less interested in one to whom they assign little power. The behavior of the group, in turn, will usually be interpreted by the person who is the target of such actions as an indication of his value to the group since he is not aware of the power being attributed to him by the members. (If he had known about his power, this may have caused him to perceive and evaluate the group's behavior differently.) The exercise vividly illustrates for the students

that their own behavior made one person feel that he was given attention and was appreciated, while the other felt ignored and unwanted—all of this because of the members' direct reaction to the power attributed to the newcomers.

The demonstration of these phenomena in the classroom provides an opportunity to experience, examine, and discuss forces at work in hierarchical groups which otherwise remain at a verbal level of awareness.

References

- HURWITZ, J., ZANDER, A., & HYMOVITCH, B., Some effects of power on the relations among group members. In D. Cartwright & A. Zander (Eds.), *Group dynamics, research and theory*. Evanston, Ill.: Row, Peterson, 1953. Pp. 483-492.
- JACKSON, J., Analysis of interpersonal relations in a formal organization. Unpublished doctor's dissertation, Univer. of Michigan, 1952.
- KELLEY, H., Communication in experimentally created hierarchies. *Hum. Relat.*, 1951, 4, 39-56.
- LIPPITT, R., POLANSKY, N., REDL, F., & ROSEN, S., The dynamics of power. *Hum. Relat.*, 1952, 5, 37-64.
- PEPITONE, A., Motivational effects in social perception. *Hum. Relat.*, 1950, 3, 57-76.
- THIBAUT, J., An experimental study of the cohesiveness of underprivileged groups. *Hum. Relat.*, 1950, 3, 251-278.

[CHAPTER 10] *The Measurement of Learning: To Count Is To Number*

Introduction

Throughout this book of readings we have considered the purpose of the schools and teachers to be the promotion of desirable behavior changes. The teacher's first task is to determine what response he wants his students to make. We have indicated that, unless the teacher is clear about the behavioral objectives he wishes the students to achieve, he has no sound basis for selecting particular learning materials, procedures, or activities. We have stated that the objectives which the teacher can select are limited by social and community pressures on the school (see Introduction, p. 527), but the educational objectives of the community are rarely defined as specific behavior changes. For example, the parent tells the teacher that he wants his child "to know how to read." But "knowing how to read" comprises many behavioral responses. For example, the child must be able to (1) recognize that different marks on the printed page are letters; (2) connect these letters with sounds he has heard; (3) combine them into syllables and words; (4) recognize words by their general shape; (5) connect a word he has read with the same word he has heard; (6) relate words to meanings; (7) discriminate between related words and meanings; (8) pronounce the words he reads—to read aloud; (9) follow directions on the printed page; (10) combine syllables into sounds for new words he has never seen or heard, etc. "Knowing how to read" is not a specific enough objective to guide the efforts of either the child or his teacher. How-

ever, specific objectives narrow the selections of appropriate materials and procedures.

The articles in this book have been concerned with the conditions of practical learning situations. For example, there have been discussions of motivation, reinforcement, concept formation, principle learning, transfer. We have also discussed various characteristics of learners, such as ability and modes of communication. In general we have discussed what we know from experimental psychology and the psychology of individual differences, and what we need to know from educational psychology in order to make practical learning situations more efficient. By efficient learning we have meant learning that results in the behavior changes which have been previously designated by the teacher as desirable and which will be preserved over long periods of time for later use by the student. It has been assumed that a greater knowledge of the characteristics of the learning situation will furnish better guidelines for the management of educational activities and increase the likelihood that efficient learning will occur.

This last chapter is devoted to the measurement of learning, that is, testing. We shall take the position of Adkins (pp. 576-586) that testing is an integral part of the educative process. By testing, the teacher tries to find out if the student's efforts to make the designated change in behavior, as well as his own efforts to engineer that change, have been successful. Where they have not been successful it is incumbent upon both teacher and student to take remedial action, perhaps trying quite different procedures for achieving the same change. One cannot consider the teaching or learning process complete until all three of the following requirements have been provided for: (1) selection of the behavioral objectives, (2) selection of appropriate materials and procedures, and (3) testing.

Tidy as this model is, students know from their own experiences that it is only roughly approximated in the schools. We cannot discuss here the carelessness or superficiality with which objectives are selected and defined, or how frequently the objectives are ignored after they have been selected. Nor can we discuss here how haphazardly learning procedures and materials are selected. Testing has received no better, and sometimes even worse, treatment. Teachers frequently see test construction and scoring as an excruciating

burden; to them it is a ritualistic practice which must be observed, like erasing the boards, but which is essentially divorced from the serious business of teaching and learning. Students see tests as great trials. Neither teacher nor student seems to understand that unless some careful and constant effort is made to find out whether or not one is learning what one is supposed to be learning, everyone operates in the dark. If a manufacturer, after erecting buildings, purchasing machinery, hiring workers and supervisors of workers, and manufacturing a product, were then to show no interest in the product, we would have a situation similar to that which exists in testing. But in education it is perhaps not a lack of interest that makes us careless or hesitant in finding out how successful or unsuccessful we have been: there may be vague but persistent misgivings about the uses of the test scores.

Not only is testing important for continued learning, it is equally important in the determination of the student's future educational and social advancement. The tests which the student takes, especially in high school and college, are the chief bases for determining his grades. The grades are the basis for many future decisions which affect him later. His parents base their decisions on how much money to provide for his future education on grades. The college bases its decision on whether or not to admit him partly on his grades. His advancement within the school and college is determined by the grades he receives. Employers, who regard education with more favor now than when they were students, base decisions for hiring on grades. All of these individuals are assuming that the information provided by the grades is some indication of how well the student will succeed in some future endeavor. If this information is misleading, there can be unpleasant consequences for both the student and the community. Improved testing can furnish more reliable information upon which to base grades.

Relationship of Readings in Chapter 10

The excerpt from Bloom's *Taxonomy* points up the need for education objectives and considers the range of intellectual objectives possible. Adkins discusses the need for tests and measurement as a means of upgrading the present educational system. Ebel shows how

we can improve our tests by careful construction and analysis of them. A second, more controversial article by Ebel raises the question of test validity. The final article is a reply to Ebel from Littell, who believes that testing must have more to do with an empirical "reality" and less to do with other tests.

BENJAMIN BLOOM, Editor
University of Chicago

Educational Objectives and Curriculum Development*

In the introduction to this chapter the importance of establishing behavioral objectives was discussed. The educative process cannot be efficient unless it is directed toward the achievement of specific tasks. "Exactly what is it that I want the students to learn?" must be the unavoidable first question of the teacher. On closer inspection, however, the teacher often discovers that educational tasks are complex and that it is necessary to reduce them to subtasks arranged in a sequence conducive to efficient learning. Perhaps the most beneficial effect of programmed learning, as discussed in Chapter 3, is that it forces the teacher-programmer to make a detailed analysis of the components of complex learning tasks and to experiment in finding out what the best order of presentation might be.

The following reading discusses educational objectives—their sources and classification. The *Taxonomy* is a classificatory system designed for use in constructing tests that would measure significant cognitive change. Although the *Taxonomy* recognizes the importance of simple remembering or the recall of information as an initial educational objective, its chief con-

* Reprinted from *Taxonomy of Educational Objectives*, Benjamin Bloom, ed., Longmans, Green, New York, 1956, pp. 25–31. Courtesy of David McKay Company, Inc.

cern is the classification of intellectual abilities and skills beyond remembering. In general, its particular concern is with the student's ability to use the information he has gained in complex problem solving.

The reading can be compared with those of Adkins (pp. 575-586) and Ebel (pp. 586-597): (1) How could the testing program described by Adkins fruitfully employ the *Taxonomy*? (2) How does Ebel's scheme for classifying text items compare with that of the *Taxonomy*? (3) What intellectual abilities and skills (beyond remembering) would you specify in a subject or skill area you plan to teach?

We have had some question about the relevance of this section in a handbook devoted to the details of a classification system. We have finally included it because we believe the classification and evaluation of educational objectives must be considered as a part of the total process of curriculum development. Some of these considerations help to clarify the distinctions made in the taxonomy. It is hoped that many teachers will find this chapter useful as a summary of some of the arguments for inclusion of a greater range of educational objectives than is typical at the secondary school or college level.

Problems of developing curriculum and instruction are usually considered in relation to four major types of questions.

1. What educational purposes or objectives should the school or course seek to attain?
2. What learning experiences can be provided that are likely to bring about the attainment of these purposes?
3. How can these learning experiences be effectively organized to help provide continuity and sequence for the learner and to help him in integrating what might otherwise appear as isolated learning experiences?
4. How can the effectiveness of learning experiences be evaluated by the use of tests and other systematic evidence-gathering procedures?

We are here concerned primarily with the first of these questions: the formulation and classification of educational objectives.

By educational objectives, we mean explicit formulations of the ways in which students are expected to be changed by the educative process. That is, the ways in which they will change in their thinking, their feelings, and their actions. There are many possible changes that can take place in students as a result of learning experiences, but since the time and resources of the school are limited, only a few of the possibilities can be realized. It is important that the major objectives of the school or unit of instruction be clearly identified if time and effort are not to be wasted on less important things and if the work of the school is to be guided by some plan.

The formulation of educational objectives is a matter of conscious choice on the part of the teaching staff, based on previous experience and aided by consideration of several kinds of data. The final selection and ordering of the objectives become a matter of making use of the learning theory and philosophy of education which the faculty accepts.

One type of source commonly used in thinking about objectives is the information available about the students. What is their present level of development? What are their needs? What are their interests? Another source for objectives is available from investigations of the conditions and problems of contemporary life which make demands on young people and adults and which provide opportunities for them. What are the activities that individuals are expected to perform? What are the problems they are likely to encounter? What are the opportunities they are likely to have for service and self-realization?

Another source of suggestions for objectives comes from the nature of the subject matter and the deliberations of subject-matter specialists on the contributions their subject is able to make to the education of the individual. What is the conception of the subject field? What are the types of learning which can arise from a study of that subject matter? What are the contributions that the subject can make in relation to other subjects?

It is likely that a consideration of these three sources will result in a suggested list of objectives which require more time and effort than the school has at its disposal. The problem of selecting among

possible objectives as well as the determination of relative emphasis to be given to various objectives requires the use of some guiding conceptions. The philosophy of education of the school serves as one guide, since the objectives to be finally included should be related to the school's view of the "good life for the individual in the good society." What are the important values? What is the proper relation between man and society? What are the proper relations between man and man?

Finally, educational objectives must be related to a psychology of learning. The faculty must distinguish goals that are feasible from goals that are unlikely to be attained in the time available, under the conditions which are possible, and with the group of students to be involved. The use of a psychology of learning enables the faculty to determine the appropriate placement of objectives in the learning sequence, helps them discover the learning conditions under which it is possible to attain an objective, and provides a way of determining the appropriate interrelationships among the objectives.

It should be clear from the foregoing that objectives are not only the goals toward which the curriculum is shaped and toward which instruction is guided, but they are also the goals that provide the detailed specification for the construction and use of evaluative techniques.

A test of the achievement of students is a test of the extent to which the students have attained these educational objectives. An achievement test is an adequate and valid test if it provides evidence of the extent to which students are attaining each of the major objectives of the unit of instruction.

The cognitive objectives derived from a process like that described in the foregoing paragraphs may, for discussion purposes, be divided into two parts. One would be the simple behavior of remembering or recalling knowledge and the other, the more complex behaviors of the abilities and skills. The following section discusses these two divisions in turn, considering their nature, their appearance in the taxonomy, and their place in the curriculum.

KNOWLEDGE AS A TAXONOMY CATEGORY

Probably the most common educational objective in American education is the acquisition of knowledge or information. That is,

it is desired that as the result of completing an educational unit, the student will be changed with respect to the amount and kind of knowledge he possesses. Frequently knowledge is the primary, sometimes almost the sole kind of, educational objective in a curriculum. In almost every course it is an important or basic one. By knowledge, we mean that the student can give evidence that he remembers, either by recalling or by recognizing, some idea or phenomenon with which he has had experience in the educational process. For our taxonomy purposes, we are defining knowledge as little more than the remembering of the idea or phenomenon in a form very close to that in which it was originally encountered.

This type of objective emphasizes most the psychological processes of remembering. Knowledge may also involve the more complex processes of relating and judging, since it is almost impossible to present an individual with a knowledge problem which includes exactly the same stimuli, signals, or cues as were present in the original learning situation. Thus, any test situation involving knowledge requires some organization and reorganization of the problem to furnish the appropriate signals and cues linking it to the knowledge the individual possesses. It may be helpful in this case to think of knowledge as something filed or stored in the mind. The task for the individual in each knowledge test situation is to find the appropriate signals and cues in the problem which will most effectively bring out whatever knowledge is filed or stored. For instance, almost everyone has had the experience of being unable to answer a question involving recall when the question is stated in one form, and then having little difficulty in remembering the necessary information when the question is restated in another form. This is well illustrated by John Dewey's story in which he asked a class, "What would you find if you dug a hole in the earth?" Getting no response, he repeated the question; again he obtained nothing but silence. The teacher chided Dr. Dewey, "You're asking the wrong question." Turning to the class, she asked, "What is the state of the center of the earth?" The class replied in unison, "Ig-neous fusion."

John Dewey's story also illustrates the rote recall nature of some knowledge learning. The emphasis on knowledge as involving little more than remembering or recall distinguishes it from those conceptions of knowledge which involve "understanding," "insight," or

which are phrased as "really know," or "true knowledge." In these latter conceptions it is implicitly assumed that knowledge is of little value if it cannot be utilized in new situations or in a form very different from that in which it was originally encountered. The denotations of these latter concepts would usually be close to what have been defined as "abilities and skills" in the taxonomy.

Whether or not one accepts this latter position, it is sufficient to note that knowledge by itself is one of the most common educational objectives. The most cursory reading of the standardized tests available or of teacher-made tests would indicate that tremendous emphasis is given in our schools to this kind of remembering or recall. A comprehensive taxonomy of educational objectives must, in our opinion, include all the educational objectives represented in American education without making judgments about their value, meaningfulness, or appropriateness. Knowledge, therefore, is one of our taxonomy categories.

The knowledge category in particular and, as noted earlier, the classifications of the taxonomy in general range from the simple to the more complex behaviors and from the concrete or tangible to the abstract or intangible. By simple we mean elemental, isolable bits of phenomena or information, e.g., "the capital of Illinois is Springfield," or "Arkansas contains much bauxite." Thus, our base subclassification is titled "knowledge of specifics." At the upper end of the knowledge category the subclassifications refer to more complex phenomena. Thus, remembering a theory is a more complex task than remembering a specific such as the capital of a state. Knowledge of the theory of evolution, for instance, would be very complex. Accordingly, the subclassification at the complex end of the knowledge category is titled the "knowledge of theories and structures."

The knowledge categories may also be viewed as running from concrete to abstract. Thus, in general, knowledge of specifics will refer to concrete, tangible phenomena: "Insects have six legs"; "Most glass is brittle." But the more complex categories, as, for example, the name "knowledge of theories and structures" implies, tend to deal with abstract phenomena.

It might sometimes be useful for taxonomy purposes to distinguish knowledge with regard to the different specialties, fields of

knowledge, or subdivisions of work in our schools. Thus, it would be possible to distinguish knowledge about the social sciences from knowledge about the physical sciences, and knowledge of physics from knowledge of chemistry, etc. Likewise, knowledge about man could be distinguished from knowledge about physical objects, etc. The taxonomy as developed here should be applicable to any of the subdivisions of knowledge or educational units in which school curricula are divided, but no attempt will be made to make all the possible applications or subdivisions in this Handbook. The reader may wish to develop such further classifications as are necessary for his work, using the taxonomy as a basis.

DOROTHY C. ADKINS
The University of North Carolina

Measurement in Relation to the Educational Process *

Most teachers know that they must construct and give tests (if for no other reason than that it has always been done), but few know why. Students also expect to be tested, but even fewer students than teachers know why. Of course, both teachers and students know that course grades are largely based on tests, but no one has ever really explained grades.

In the following article, the author attempts to explain what a test is and why we need it. The theory of testing proposed here is based on Skinner's theory of learning (pp. 138-139). This weaving together of testing and learning is particularly appropriate for an anthology on human learning in the school. In reviewing this article, the student may find the following

* Reprinted and abridged with the permission of the author and the publisher from the article of the same title, *Educational and Psychological Measurement*, 18 (1958), 221-240.

suggestions helpful: (1) One could almost say that Adkins believes that teaching is testing. How does her description of the "cycle" of the educative process support this view? (2) She has stated that Skinner may become the "sponsor of the century's most massive measurement movement" (p. 584). What is she referring to and how does this relate to her cycle theory of the educative process? (3) What does she believe is the chief purpose of achievement tests? How is this related to transfer of training? (4) What relationship is there between testing, instruction, and individual differences? (5) How would Adkins build tests directly into the curriculum? Consequently, how would she report grades? What advantages would this have over the present system?

The measurement of change in the learner—all too often informal, subjective, and unreliable—is an inescapable aspect of various phases of the cycle of activity comprising the educational process. For purposes of clarification, it may be separated into the following steps:

1. Defining behavioral goals or objectives.
2. Planning stimuli (that is, curricular materials) and teaching methods.
3. Predicting the degree to which the objectives can be attained.
4. Applying the teaching methods to the selected curricular content or adjusting the goals or the content and methods.
5. Testing achievement.
6. Defining new behavioral objectives in the light of how well the old ones were met, and so on.

This succession has been called a cycle because it has no boundaries. Each new series must be adjusted upon the basis of the results of the preceding ones. Throughout the sequence, evaluation of either current status of the learner or of change in him must be more or less continuous, whether or not the appraisal is recognized as such or is conducted by means of formal tests.

Returning to the first step, the behavioral goals defined cannot be realistic unless they are based upon a knowledge of the abilities of the prospective learners to reach alternative ends that might be set. At once the necessity of measuring both various kinds of aptitude and previous achievement becomes apparent.

The second phase of the cycle, planning the curricular content and teaching methods, must be related to previously acquired knowledge or skills of the learners and to the readiness with which they can acquire the contemplated new abilities. Again, the decisions need to take into account the learners' present status. Likewise, teaching techniques should be modified in view of the capacities of the learners.

The prediction of whether or not the goals are realistic, listed as the third step, requires consideration of the aims themselves, the selected curricular content, and the proposed teaching techniques in relation to the previous attainments and aptitudes of the learners. Will a particular learner, presented with certain stimuli with specified amounts of reading, explanation, drill, recitation, dosage, and so on, demonstrate the desired behavioral change? Or will his time and that of the teacher largely be dissipated? If a prediction is to be made—as indeed it should be—appraisal of the learners' abilities again must be made.

The fourth phase, the application of particular teaching methods to selected curricular content, can be successful only if the teacher is vigilant in measuring progress toward the desired objectives. If retardation is evident, then teaching technique may be altered, content supplemented, or the goal adjusted. The teacher also should be responsive to more rapid accomplishment of the objective than was anticipated, else otherwise excellent learners will shy away from scholarship through sheer boredom. Here assessment of change in the learner as it is occurring must be prominent.

The fifth step calls for a test of the attainment of the objective of a unit of instruction. Such a test could be regarded as one of current status. Whether or not the achievement reflected in the test can be attributed to the unit of the educational process in question or to the particular teaching involved can not be resolved by a single administration of a test, however. Rather, test results must be interpretable in terms of how much change they reflect. Two com-

parable tests, given before and after some particular experience, such as exposure to a segment of content by means of certain methods, often will permit an inference as to the extent to which variation in behavior has taken place. Sometimes a conclusion is warranted that such change as is evident is associated with the intervening events, especially if a control group or some other appropriate experimental design has been employed.

The educational process should be dynamic. It should not concentrate fixedly upon an unchanging ultimate goal the attainment of which eventually may be tested. Rather, a series of proximate or intermediate aims should be established, with each subject to constant readjustment in the light of frequent appraisals of the attainment of previous ones. Viewed in this light, the setting of educational objectives can not be divorced from the appraisal of progress toward them. . . .

The ultimate purpose of an achievement test, like that of an aptitude test, is to predict future behavior. Seen in proper perspective, an achievement test must serve this function if it is to go beyond the narrow limits of today's activities. Indeed, were educators completely unable to predict future performance from present achievement, the process of education would become chaotic. Any basis for selecting curricular content, mastery of which presumably would have positive transfer to later situations, would disappear. The inescapable fact, then, is that we are interested in present achievement primarily because it permits a prediction of future performance, however informal and often neglected that prognosis is.

Frank acknowledgment of this fundamental purpose of educational achievement tests should lead to revisions in the measuring devices themselves, in the curriculum, and in teaching methods. Recall with me a study reported by Sidney Pressey in *Psychology and the New Education* in 1933. Students' scores on end-of-course examinations in various high-school and college subjects were compared with their performance on the same examinations after one and two years. To quote, "These students showed a knowledge of about three-quarters of the material covered in the test at the end of the course, less than half after one year and about one-quarter

after two years." One hesitates to project these loss of retention curves to even one more year! Pressey himself speculates as to whether solace may be found in the thought that students retain general understanding, principles, points of view, methods of attack, rather than minimally essential facts. Although too many educational achievement tests are limited largely to facts acquired by rote memorization, Pressey concludes by arguing persuasively that we are but little better off in the higher echelons of the thought processes. Finally, he suggests that the slight permanence of much school learning is chargeable to the curriculum, which evidently is not based upon content needed or used outside the classroom.

Yet, while the foregoing argument may hold for, say, algebra, which may appear to fade through disuse, it can scarcely apply to the basic skills of reading and writing, for example, which are practiced quite regularly and in which college students also reveal little permanence of learning.

A repetition of such an experiment in almost any field would yield equally sobering results. The basic difficulty may lie in the fact that materials learned to a level below mastery disappear from our repertoires for the very reason that they are not fully mastered, so that correct responses are not made often enough to get sufficient reinforcement in the normal course of events to perpetuate them. Surprisingly enough, although the advantageous effects of strategic degrees of overlearning upon relearning and eventual permanent mastery have been common knowledge to educators since the 1885 publication of Ebbinghaus' work, teaching practice reflects meager appreciation of them.

The suggested solution will lie in teaching each learner to the mastery level those materials that he is capable of really mastering, by suitable teaching methods and with more or less continuous progress appraisals based upon tests of defined educational objectives. The remedy is not simply to multiply the use of tests. More drastic surgery is indicated. Tests need to be integrated into the entire educational process, as heretofore suggested. To reiterate briefly, we must assess the present abilities of the learners, prepare curricular materials appropriate to the ability levels, adapt teaching methods to the learners and the content to be learned, establish appropriate degrees of overlearning by empirical means, and constantly appraise the learning as it is taking place. . . .

The educational process is expensive, to the society that supplies the dollars and to the learner who contributes the time. As teachers' salaries rise and as the number of pupils increases, the total cost becomes larger. Despite the fact that I think teachers are grossly underpaid, I would at the same time point out that our educational system demands radical improvement. Consider, if you will, the hundreds of hours that are devoted to instruction in English by the time a person becomes a college senior. Then ponder his inability to write. Turn to arithmetic, the time devoted to fruitless drill, and later to algebra and geometry. Then reflect upon the ineptitude of a typical college student in solving the most elementary equation or even in adding a column of figures. Recall the idiosyncrasies in spelling that confront you in personal correspondence. Contemplate the current reading hubbub, by no means without foundation. Shudder at the prevalent superstitions and gullibility of the American public.

This does not mean that we should return to the "good old days," nor that as a people we are worse off than our immediate forbears as far as reading, writing, and arithmetic are concerned. More of us are about equally bad off, and the total effect is thus more devastating. A large part of what transpires under the name of education should be regarded candidly as serving primarily the function of custodial care of the callow. (It also, of course, provides an outlet for members of the teaching profession. And it slows down traffic on school days.) This safekeeping of the young is by no means a thankless endeavor, because it frees parents for several hours a day for other pursuits and hence presumably contributes to their tranquility. Thus the national trend toward starting children to school at earlier ages and keeping them there through early adulthood is not unrelated to the relief of parents from some of the responsibility for their care. As far as progress toward educational objectives is concerned, however, no one believes that schools operate at anywhere near 100 per cent efficiency.

This is not a criticism of individual teachers nor of the profession of teaching. It is an indictment of an educational system within which teachers operate and which is so firmly entrenched by long tradition that no single teacher or group of teachers or school can

hope to resist it. Indeed, widespread and drastic educational change is warranted.

Consider a more or less typical fifth-grade group. With a mean age of around 10, it may contain some as young as 8 and others who have repeated a grade or more and who are thus as old as 13, or even older. Their general intelligence, in terms of the rather coarse I.Q. measure, will vary from perhaps 55 to as high as 170 or more. Their mental ages (equally crude, of course) might range from around 7 to 15 or so. If the school is large enough, some ability grouping doubtless would be used to reduce these ranges. Even so, they are typically pretty large. Children's differences in the several more or less native aptitudes that go to make up the common general intelligence measure are enhanced by the fact that they have acquired widely varying degrees of competence in the many skills and knowledges to which they have been exposed. So what happens? With notable exceptions, every child is given identical assignments, exposed to essentially the same teaching methods with the same amount of learning time, and tested occasionally by identical tests. Teachers are well aware that the amounts learned under such conditions will vary markedly, and achievement tests commonly reveal wide ranges of ability. So inured are we to these differences that their very presence provides a clue to the validity of the tests. When the majority of scores cluster about a particular point, we say that the test is not discriminating among individuals, leading to the conclusion that it is not a valid measuring device. If all learners do well on a test, we plan to replace it by one more difficult in order to reflect the different degrees of mastery of subject matter that our educational system insures will exist.

Under such a plan, the educationally poor become more impoverished, but the rich are not strengthened. By the nature of the system, the weaker student is forced to endure failure over and over again. Aside from the questionable desirability of such negative motivation, the blighting effects of constant frustration upon personality development should be a matter of grave concern. The abler student, on the other hand, often experiences success too readily, with no genuine effort. This painless achievement again may have harmful outcomes when the student later encounters situations that require exertion.

Along what lines lies a solution, short of providing a private tutor for every child? The first major need is for extensive curriculum revision, in the direction of a large number of learning units scaled according to difficulty within subject-matter areas. Each such area should be limited to a body of knowledges or set of related skills that require the same pattern of abilities, which include both the effects of native aptitudes and previously acquired knowledges or skills. An initial step in scaling learning units would be the testing of these abilities of prospective learners. Then tentative curricular units could be tried out with groups of varying ability patterns by different methods of presentation. Doubtless experienced teachers could effectively reduce the number of methods to be tried by excluding in advance the ones likely to be least effective. Comprehensive subject-matter tests based upon materials taught by whatever method is current in a particular school would provide a start. Then, by some combination of expert teacher judgment and empirical test data, the units of instruction could be placed in appropriate order. The next step would be to develop mastery tests for each unit. These would be entirely different from the tests commonly now administered at the end of a unit of instruction, in which the teacher is disappointed if he can not observe wide individual differences. Rather, the learner would be expected to persist at curricular units of a given difficulty level until he had achieved a standard degree of mastery, at which time he would be ready for the next higher level.

Note some of the features of such an educational plan. The learner is not frequently assigned content entirely beyond his current ability level. Nor is he ever wasting time upon materials absurdly easy for him. He rarely experiences failures. The questionable luxury of not exerting effort is disallowed. The learner competes always at his own ability level, not being exposed to content for which he lacks prerequisites. He becomes well acquainted with his limitations as well as with his strengths. His degree of mastery of what he is learning is continuously appraised, so that at all times both he and his teacher can know exactly how he is progressing.

Does this emphasis upon positive motivation and mastery of one learning unit before proceeding to the next mean that by some magic, at the end of several years of exposure to the educational

process, we will have been able to eliminate the individual differences so conspicuous at its beginning? Not at all. If anything, the discrepancies will be more pronounced. Moreover, reporting upon such differences as will exist to prospective employers or to higher educational institutions could well be more detailed and far more precise than under the present system. In 1984, we may be able to say to an employer or to a university that John Doe has mastered mathematics unit 1728, English grammar 642, spelling 1021, physics 305, and so on. The personnel officers will by that time have conducted studies to enable them to predict from such data his relative chances of success as an outer-space radiological isotopist or as a student in a modern household engineering curriculum.

Are steps now being taken to reduce the formidable inefficiency and wastage of this mass baby-sitting movement that goes under the guise of the educational process? Undoubtedly isolated schools and rare teachers exist in contradiction to the perhaps shocking phraseology in which I have characterized modern education. But are there discernible trends, omens of significant departures from an educational system geared to presenting identical doses of pabulum to be partially digested within a uniform allotment of time? I am not to be satisfied—nor should you be—by having each teacher know the I.Q. of her every pupil, by further applications of “ability grouping,” by myriad “opportunity rooms,” or by what need no longer be called “progressive education.” What we must seek is a situation in which a sizable share of curricular materials, graded according to the ability pattern and levels required, is presented to the learner in such a way that he can work at his own rate with a record of his progress constantly available to him and to his teacher. The measurement approach must be applied in the scaling of the curricular units as well as in the appraisals of progress. A dearth of such graded learning units exists. In fact, one is hard pressed to cite any aside from Ernest Horn’s admirable early work in grading 3500 common words as to spelling difficulty and the contributions of Guy Buswell and others in elementary arithmetic.

A study aimed at the production of graded reading exercises, beginning with ones appropriate for elementary school grade 4 and continuing to the graduate school level, has recently been undertaken by Thelma G. Thurstone in the Psychometric Laboratory at

the University of North Carolina. Currently the project involves 4500 items; and once they have been scaled, additional items can readily be added. This study should represent a major advance in the teaching and evaluation of reading skills.

The same B. F. Skinner with whom I have agreed and disagreed in turn is also developing scaled curricular materials as one phase of his work on teaching machines, about which many of you will have heard. Briefly, his plan calls first for the careful preparation of such materials appropriately ordered as to difficulty. Successive elements of each learning unit are presented to the learner in a window of a machine. He is alternately given information, required to practice or apply the information, and tested on each phase until he has mastered the unit. The machine can be set in several ingenious ways, so that, for example, the learner must finally proceed through a mastery test, say, three times in succession without error. Moreover, the robot can reject questions on which the learner has already demonstrated mastery so that he need not devote time to them. Skinner himself, as I interpret him, has misgivings about the reactions of both teachers and testers to his inventive approach to the educational process. He seems apprehensive lest they fear that they will soon be out of business, the teachers because the machine will be presenting materials in their stead and the testers because somehow appraisals of mastery will no longer be needed. To the contrary, teachers as both curriculum specialists and test construction experts will be needed in large numbers, even aside from their roles as discussion leaders, motivators, socializers, stimulators of creative thought, and so on. Skinner, in spite of his possibly ephemeral scorn of aspect testing, will come to recognize that such measurement is a very integral feature of his teaching machine. If his plans fully materialize, posterity may consign his pigeons to extinction and pay homage to him chiefly as the sponsor of the century's most massive measurement movement.

Pressey experimented with a similar idea in the late twenties or early thirties. He had a machine, like a miniature cash register, by which a student could indicate answers to objective test questions and immediately learn whether his answer was right or wrong. Basically the same idea was adopted by Angell and Troyer in their Self-Scorer. It also provides a means whereby a student taking a test

continues to select answers until he obtains the right one, which is revealed by a red mark under the hole on the answer sheet that he has punched out. While we await the commercial feasibility of Skinner's latest venture, a relatively simple and inexpensive device such as this one may help to span the breach between instruction and testing. It is practicable for a small number of questions in daily discussion periods or for longer achievement tests.

On a less dramatic front, such college examiners as Paul L. Dressel at Michigan State University, working over the past several years with the instructional staff on evaluation of the objectives of college courses, has pursued an integrated approach to evaluation and teaching to the extent that they become indistinguishable from many points of view. To quote Dressel: "Evaluation does not differ from instruction in purposes, in methods, or in materials and can be differentiated from instruction only when the primary purpose is that of passing judgment on the achievement of a student at the close of a period of instruction."

The general outlines of an educational system that I have presented and the significant role that I have visualized for measurement in relation to the educational process may seem a far cry from such an institution of higher learning as, for example, Sarah Lawrence College. There, according to the jacket of *Essays in Teaching*, edited by its president, Harold Taylor, exists "a freedom of intellectual inquiry unencumbered by the apparatus of assignments and tests." This is so overwhelming that I perforce discredit it. Simply because assignments are not organized or just because they are adjusted to the individual students does not mean that no assignments are made. As for tests, from what I have heard and read about the evaluation of students' abilities, needs, and interests at Sarah Lawrence, far more examining is done there than at the average college. The disuse of end-of-course marks or of any kind of grades that resemble the familiar A, B, C's or the 70-to-100 scale does not indicate lack of measurement. The fact that much of the evaluation depends upon subjective impressions gleaned from essays and individual conferences rather than upon objective examinations does not signify that no measurement is taking place. Conceivably, such testing, based upon close personal contact between student and teacher, may be both less reliable and more valid than many of the

objective tests currently extant. As I am sure will be clear by now, however, for a country-wide educational program I would rather place my chips on a more orderly curriculum and a more highly organized plan for measurement, closely geared together at every stage of the educational process.

ROBERT L. EBEL
Educational Testing Service
Princeton, New Jersey

Procedures for the Analysis of Classroom Tests *

The question posed by Ebel in this article is "How good are your objective tests?" That objective tests have often possessed only degrees of "badness" can hardly be denied. Frequently they require little more than the recognition of a correct factual answer of one or two words in response to a rather simple-minded question of information. Sometimes the test items are so carelessly constructed that they unwittingly reveal more information than they ask for, and in every case, the items give the correct answer. Students may properly boast about the high degree of test sophistication that they have developed through years of subjection to bad tests. It is even conceivable that in some universities one could obtain a degree by simply "unfoiling" foils on poorly constructed multiple-choice tests. True-false tests have been so poorly constructed that the students who know the most often obtain the lowest scores.

* Reprinted and abridged with the permission of the author and the publisher from the article of the same title, *Educational and Psychological Measurement*, 14 (1954), 352-364.

The solution seems to be better tests—of both the essay and the objective type. The following article is limited to a consideration of the latter. In reviewing this article the student may find these questions helpful: (1) What does Ebel mean by relevance? What type of validity would Adkins call this? (2) Ebel has given examples of test items which range from simple recall of factual information to “application” items. Of what characteristic of practical learning situations, frequently referred to in these articles, would such items help to measure the successful outcome? Try to write an “application” item in a subject you plan to teach. (3) What does Ebel mean by discriminating power? If the teacher builds a test using this concept, how would this test differ from many others you have taken?

Defining the Problem

It is generally agreed that classroom tests at the college level, and at most other levels as well, are in need of improvement. Poor tests are the most frequent target of criticism when students rate their instructors and courses. Instructors themselves sometimes admit dissatisfaction with their tests and ask for suggestions as to how they may be improved. But the clearest indications of weakness in typical classroom tests are found in critical and statistical analyses such as those described in this paper.

When a central examinations service was set up at the University of Iowa a decade ago the principal purpose specified for it was to aid instructors to improve their classroom tests. They were relieved of the clerical burdens of duplicating and scoring objective tests so that they might have more time to write better items. Consultation service was offered on testing problems. Files of sample tests were accumulated. Statistical computations, and especially item analyses, were provided. Any or all of these services were offered on request and without charge to instructors or departments. No attempt was made to require instructors to use these services, or to change their examination practices. Under these permissive conditions, which still seem desirable, it has become apparent that test analysis is the

most powerful tool for making progressive improvement in classroom tests.

Originally the statistical service offered consisted of calculation of percentile ranks and analysis of items for difficulty and discrimination. While this information was useful, it did not give any simple or reasonably complete answer to the instructor's basic question, "How good is this test?" In an effort to answer the question more adequately, as well as in the hope of making instructors more familiar with the standards of good classroom testing and more clearly aware of the limitations of their tests, we have recently expanded our test analysis service.

The purpose of this paper is to describe this extended analysis and to explain the standards associated with it. No basically new techniques of test analysis have been invented. The main effort has been to apply existing techniques systematically and to develop reasonable standards of excellence for classroom tests.

For several fairly obvious reasons this report will deal only with the analysis and improvement of objective tests. This does not mean that all instructors at Iowa *do* use objective tests, for many do not. It does not mean that we believe instructors always *should* use objective tests, for there are certainly some situations in which essay tests are more effective than objective tests, though many of the reasons given for preferring essay tests seem to be based on prejudice and misconception. It does not mean that only objective tests need improvement, for, on the contrary, the essay tests which many instructors prefer to use are probably as much in need of improvement as are objective tests. To lead a college faculty to use objective tests at the right times and essay tests for the right reasons is a major problem in the improvement of examinations. But it is one which lies outside the scope of this paper.

Analysis for Relevance

The essential qualities of a good educational achievement test or test item are *relevance* and *discriminating power*. Practically all the suggestions offered for good item writing or good test construction relate directly or indirectly to these two basic qualities. When instructors are urged to base test items on course objectives rather

than on course content, to emphasize important and useful information rather than trivial or academic details, and to reward understanding and application rather than rote learning or repetition, the purpose is to improve test relevance.

When instructors are urged to select items which most well qualified examinees will answer correctly and most poorly qualified examinees miss, to express those items clearly, to make sure the correct response is demonstrably better than any alternative, and to make sure that each incorrect alternative has some basis for attracting poorly qualified examinees, the purpose is to improve the discriminating power of individual items, and thus to improve the discriminating power of the test as a whole.

Some of the important factors in relevance and discriminating power are outlined on the Test Analysis Report form which is shown as Table 1. These factors are listed in the first column under the heading *Characteristic*. Desired standards are listed in the second column, headed *Ideal*. Results obtained from the analysis are presented in the third column, headed *Observed*. Verbal evaluations of these results are presented in the fourth column, headed *Rating*. Sample data reported on this form were obtained from the analysis of a carefully developed entrance test, refined in several previous tryouts and revisions.

The analysis shows, for example, that none of the items in this test involve content details. Since such items are undesirable by our standards this aspect is rated "Good." Again, 94 per cent of the Mathematics test items involve application. While it is admittedly easier to write application items in mathematics than in some other areas, this aspect is rated "Excellent" since the desired standard is something more than 20 per cent for this category.

Only one aspect of relevance is covered in this analysis. No attempt is made to indicate the topical coverage and emphasis. All items in a given category are treated as if they were equally relevant or irrelevant when they actually are not. While such other analyses of relevance are most important, they can be made adequately only by an expert in the field covered by the test. The type of analysis reported here can be made by a test specialist, can be applied to most classroom tests, and is by no means insignificant. Over-emphasis on content details, specialized vocabulary and specific factual in-

TABLE I

University Examinations Service

State University of Iowa

TEST ANALYSIS REPORT

Test Title Mathematics Skills form 8 Number of Items 60
 Examinees L. W. Freshman Date of Test Sept. 1951

Characteristic	Ideal	Observed	Rating
I Relevance			
A. Irrelevant			
1. Content details	0%	<u>0%</u>	<u>Good</u>
2. Non-functional	0%	<u>0%</u>	<u>Good</u>
B. Relevant			
1. Information			
a. Vocabulary	<20%	<u>0%</u>	<u>Good</u>
b. Facts	<20%	<u>3%</u>	<u>Good</u>
c. Generalizations	>20%	<u>0%</u>	<u>OK</u>
2. Understanding	>20%	<u>3%</u>	<u>OK</u>
3. Application	>20%	<u>94%</u>	<u>Excellent</u>
II Discrimination			
A. Item			
1. High (more than .4D)	>50%	<u>84%</u>	<u>Excellent</u>
2. Moderate (more than .2D, up to .4D)	<40%	<u>13%</u>	<u>OK</u>
3. Low (more than zero, up to .2D)	<10%	<u>3%</u>	<u>Good</u>
4. Zero or negative	0%	<u>0%</u>	<u>Good</u>
B. Score			
1. Mean	M = 30	<u>27</u>	<u>Low</u>
2. Standard deviation	> S = 15	<u>15</u>	<u>Good</u>
3. Reliability	> .90	<u>94</u>	<u>Good</u>
4. Probable error	-	<u>2.4</u>	<u>-</u>

D = Maximum difference (number in either extreme group).

M = Midpoint of range between highest possible and expected chance score.

S = One-fourth of range between highest possible and expected chance score.

formation, with corresponding neglect of generalizations, understanding, and application, have been observed so frequently that it seems eminently worthwhile to stress this aspect of relevance.

The categories used in the analysis of relevance from this point of view are defined in general terms as follows:

1. A *content detail* item deals chiefly with instructional materials rather than with the objectives of instruction, or has no significance outside a particular classroom, or can be answered "correctly" only in terms of a particular source of information.

2. A *non-functional* item is one which is ambiguous or which can be answered correctly on the basis of subtle context clues.

3. A *vocabulary* item is one in which the principal basis for correct response is knowledge of the meaning of a single term.

4. A *fact* item is based upon some specific observation or some statement which is restricted in application.

5. A *generalization* item is based upon the summation of an extensive group of objects, events, observations or experiments, or deals with a conclusion, a principle, a trend, or a general condition.

6. An *understanding* item calls for explanation of a condition or an action, for interpretation of a statement, for knowledge of purposes or determining factors.

7. An *application* item calls for originality on the part of the examinee in dealing with a specific situation, solving a problem, recommending a procedure, or making a judgment, an evaluation, or a prediction.

The following items illustrate each category of relevance. These items were taken from a classroom test in psychology. They were selected to show clearly the differences between the categories. Not all items can be so easily classified as these. Some will appear to be "borderline cases." Others will have characteristics which do not fit any, or fit more than one category. Deliberate judgment is required in these cases, and some "errors" are likely to be made. But it is not likely that enough errors will be made to distort seriously the overall emphasis of the test.

1. CONTINENTAL DETAIL

In an experiment in which anti-aircraft gunners were taught to estimate the distance of approaching planes, it was found that

- (1) the responses were made on the basis of interrelationships among several stimuli

- (2) gunners could learn to estimate range accurately after a 24-hour period
- (3) perceptual reactions depended solely upon the size of the object in the gunsight
- (4) the gunners had a tendency to hold their fire until an approaching plane was too close

2. NON-FUNCTIONAL

In order for regression to occur, the behavior patterns that are re-adopted must have been

- (1) emotional in character
- (2) operative some time in the past
- (3) learned in an emotional situation
- (4) a more successful solution of the problem

3. VOCABULARY

Extreme attempts to overcome feelings of inferiority are the main characteristics of

- (1) aggression
- (2) regression
- (3) over-compensation
- (4) compensation

4. FACT

Aphasia is usually associated with damage of

- (1) the right hemisphere
- (2) the left hemisphere
- (3) prefrontal lobes
- (4) frontal lobes

5. GENERALIZATION

Studies of personality as a function of birth order have shown that

- (1) personality is not related to birth order
- (2) the oldest child is usually most independent

- (3) the youngest child is usually spoiled
- (4) those children who are neither oldest nor youngest are best adjusted

6. UNDERSTANDING

Scientists usually work in the artificial environment of the laboratory because

- (1) measurements can be made more precisely in the laboratory
- (2) statements of truth cannot be made unless they have been verified in the laboratory
- (3) events in nature never repeat themselves, and an event must be repeated to be studied
- (4) sources of variation in the phenomena studied can be controlled in the laboratory

7. APPLICATION

If a student had at his disposal two hours in which to memorize long lists of dates and foreign words, it would probably be most advantageous for him to

- (1) spend the whole two hours in a single unit of concentrated study
- (2) divide his study into units of about 5 minutes each
- (3) divide his study into two units of one hour each
- (4) study in units of from 20 to 30 minutes each

The percentages given as ideal for each relevance category are somewhat arbitrary, and are intended mainly to suggest the desired direction of emphasis. More weight is given to generalization, understanding, and application than to facts, terms or content details. This emphasis does not imply that content details, technical vocabulary, and specific facts are unessential to learning. It does imply that these things are means rather than ends, and that good educational achievement tests should be concerned as directly as possible with the ends, or ultimate objectives, of instruction.

The fact that these categories of relevance are not mutually exclusive is also worth noting. As arranged in the outline they consti-

tute a crude hierarchy in which each succeeding category (except the non-functional) encompasses much of the preceding category but goes somewhat beyond it. Thus, application is based on, but includes more than understanding. Generalizations are based on facts but involve something more in the way of synthesis. And all of these require some knowledge of terms and some acquaintance with content. Hence, when a test emphasizes generalizations, understandings, and applications it is also measuring the other categories indirectly. But a test which stresses content details, terms, and facts is far less likely to reveal knowledge of generalizations, possession of understanding, or ability to make applications.

Analysis for Discriminating Power

For many years our item analysis was based upon complete response counts for upper and lower achievement groups made up of the 27 per cent extremes of the distribution of criterion scores. From these counts percentages of correct response were obtained as indices of item difficulty and correlation coefficients were obtained from Flanagan's table as indices of discrimination. We recommended selection of items in the mid-range of difficulty (40 per cent to 70 per cent) and as high as possible in discrimination (above .30 if possible). Instructors were told that these policies would result in selection of items which do the best possible job of discrimination.

In time it became apparent that a combined index of difficulty and discrimination could be obtained simply by subtracting the number of correct responses in the lower criterion group from the number of correct responses in the upper. An article by Johnson indicated that while this was not a new idea it had considerable merit. Instead of expressing this upper-lower difference for each item as a proportion of the maximum possible difference as Johnson suggests, we chose to work directly with the "raw" differences. This procedure is somewhat simpler and, in our application, involves no disadvantages.

The standards suggested for this index are again somewhat arbitrary. The maximum possible upper-lower difference is equal to the number in the upper group. Items which yield a difference which is .4 or more of this maximum are called high in discrimination.

Those whose difference is less than .2 of the maximum are called low. Experience seems to indicate that these standards are high enough to suggest the need for improvement in most tests, but not so high as to make the standard appear unattainable.

Not all test constructors feel comfortable operating without the guidance of a separate index of item difficulty, especially when they use an index of discrimination such as this which favors items of 50 per cent difficulty. While they agree that items of zero or 100 per cent difficulty are useless, and those near zero or 100 per cent nearly useless, they feel that some distribution of item difficulties between these extremes is desirable. The theory is that difficult items are needed to test the good students, and easy items to test the poor ones.

This theory assumes high item intercorrelation, which is seldom if ever observed among educational achievement test items. Where item intercorrelations are low, selection of items whose difficulty is near 50 per cent tends to flatten the score distribution, to increase the dispersion of scores and thus to improve the discriminating power of the test as a whole. Selection of items to give a range of difficulties, on the other hand, tends to peak the score distribution, decrease the dispersion and thus impair the discriminating power, provided item intercorrelations are low. If item intercorrelations were high, distributed difficulty values would give better discrimination than concentration at the 50 per cent level. However, we have yet to see a classroom test in which item intercorrelations were high enough and the concentration of item difficulties about the 50 per cent level close enough even to approach a rectangular distribution, to say nothing of going further to yield the undesirable bimodal distribution which is theoretically possible. Until that happens, item difficulties per se may be safely, even advantageously, disregarded.

Analysis of Score Discrimination

The importance of dispersion in the test score distribution has already been mentioned. The amount of dispersion possible depends partly upon the location of the mean along the score scale. Its location also indicates whether the test as a whole is too easy, too

difficult, or about right. If the useful score range on a test is considered to extend from the highest possible score to the expected chance score, the best location for the mean is at the midpoint of that useful score range. For the mathematics test whose analysis is presented in detail in Table 1, the highest possible score is 60, since there are 60 items, scored 1 if correctly answered. The expected chance score is zero, since scores on this test were corrected for guessing. Hence the ideal location for the mean is a score of 30. The observed mean is 27, not far from ideal, but indicating that the test as a whole is somewhat too difficult.

The choice of a minimum standard for score dispersion was again somewhat arbitrary. When 100 or fewer scores are normally distributed, the range is approximately five times the standard deviation. For rectangular distributions the range is about 3.5 times the standard deviation. Since it appears that rectangularly distributed scores, though seldom attainable, would provide better over-all discrimination than would normally distributed scores, the compromise between 5 and 3.5 was weighted in favor of the rectangular distribution, and set at 4. That is, the ideal standard deviation should be at least one-fourth of the useful score range. For the mathematics test of Table 1 this ideal minimum value is one-fourth of 60, or 15. The observed value just reaches this standard.

The best single measure of the discriminating power of a test is its reliability coefficient. Experience indicates that carefully built tests of educational achievement can have reliability coefficients of .90 or higher. Experience also indicates that few informal classroom tests are this reliable. Hence .90 seemed a reasonable minimum standard. The odd-even reliability coefficient for the mathematics test of Table 1 is .94, surpassing the standard by a fair margin.

The principal reason for reporting probable score errors is to suggest to instructors that the scores on their tests are not perfectly "accurate" even when no scoring errors have been made, and to dramatize the need for more highly discriminating test scores. While a small probable error is better, other things being equal, than a large one, other things frequently are not equal. Because of this the better test may sometimes have a larger probable score error than the poorer. For this reason no standard for probable errors is suggested.

Concluding Remarks

The test analysis described in this paper is certainly not exhaustive. For most tests other aspects of relevance should be considered by the classroom teacher. For some tests a posteriori analyses of relevance (statistical validity studies) are possible and should be made. There probably are some types of educational achievement tests for which other types of analysis would be more satisfactory. But even with these limitations, a test analysis such as we have suggested is a useful tool which promises to be effective in improving the quality of classroom tests. It is a systematic, objective approach to the difficult problem of test evaluation.

Reference

- JOHNSON, A. PEMBERTON. "Notes on a Suggested Index of Item Validity: The U-L Index." *Journal of Educational Psychology* LXII (1951), 499-504.

ROBERT L. EBEL
Educational Testing Service
Princeton, New Jersey

Must All Tests Be Valid?*

Asking whether or not a test needs to be valid, in the fields of psychology and education, is like asking whether or not the President of the United States needs to be an American. What is validity? The author of the following article considers several definitions of this term. If the question is, "How well does Johnny comprehend what he reads?" we give him several passages of different levels and types of difficulty to read, ask him

* Reprinted with the permission of the author and the American Psychological Association, from the article of the same title, *American Psychologist*, 16 (1961), 640-647.

some questions which will reveal his understanding of the passages, and then score his responses. If the question is, "How well does the student understand the concept of motivation?" we may give him some examples of motivated behavior and allow him to make inferences about causation. In terms of the educative process as it is understood here, a test is valid if it measures the behavior change the teacher tries to promote.

Of course, as the following article indicates, the question is much more complicated than this. To some extent modern psychological thought has never freed itself from its philosophical and theological origins. This is not to indicate that psychology must completely renounce the past and face only the future. However, the philosophical tradition has often made us suspicious of appearances, and it draws our attention to an "inner reality" as the real determinant of behavior. Much of motivational theory which postulates inner "needs" belongs to this category of thought. Behavior, and as it works out, tests which sample behavior, are but a paltry reflection of the true inner state of man or of his "real" behavior. Therefore an intelligence test is frowned upon because it can never touch the inner reality which is intelligence. And it appears that it will be even harder to touch the inner reality which is creativity. But intelligence and creativity are merely hypothetical constructs, and according to Ebel, they can be little else than the tests we build and use to measure them. If the results of several independent measures of intelligence agree, these instruments give us all the validity we need. Even when we define our course objectives in behavioral terms, as has been suggested by Adkins (pp. 571-572), the statement frequently looks like a test item. In fact, in the modern school we are relying more and more on test behavior, rather than on direct observation, as a basis for evaluation.

There are many other issues which Ebel raises: (1) How does Ebel's view of validity reflect his philosophical position? (2) How might Ebel's analogy between physical and mental measurement be misleading? (3) Why would he disagree with Adkin's view of predictive validity? (4) How would his concept of meaningfulness be helpful in school practice?

Validity has long been one of the major deities in the pantheon of the psychometrician. It is universally praised, but the good works done in its name are remarkably few. Test validation, in fact, is widely regarded as the least satisfactory aspect of test development. For this the blame is usually placed on the lack of good criterion measures. To assuage their guilt feelings about inadequate test validation, test constructors from time to time urge their colleagues to go to work on the criterion problem.

It is the purpose of this paper to develop an alternative explanation of the problem, and to propose an alternative solution. The basic difficulty in validating many tests arises, we believe, not from inadequate criteria but from logical and operational limitations of the concept of validity itself. We are persuaded that faster progress will be made toward better educational and psychological tests if validity is given a much more specific and restricted definition than is usually the case, and if it is no longer regarded as the supremely important quality of even mental test.

Difficulties with Validity

DEFINITIONS OF VALIDITY

There are at least four indications that all is not well with the concept of validity as applied to mental tests. The first is that test specialists tend to differ in their definitions of the concept. Gulliksen has said: "the validity of a test is the correlation of the test with some criterion." Cureton writes: "The validity of a test is an estimate of the correlation between the raw test scores and the 'true' (that is perfectly reliable) criterion scores." Lindquist suggests: "The validity of a test may be defined as the accuracy with which it measures that which it is intended to measure, or as the degree to which it approaches infallibility in measuring what it purports to measure." Edgerton says: "By 'validity' we refer to the extent to which the measuring device is useful for a given purpose." Cronbach explains: "The more fully and confidently a test can be interpreted, the greater its validity."

No exact scientist would accept such diverse statements as operationally useful definitions of the same quantitative concept. While there is obviously some conceptual similarity there also are important divergencies. The first specifies correlation with a criterion. The second requires estimation of a corrected correlation coefficient. The third avoids statistical terms, stressing accuracy in relation to the user's intent. The fourth makes validity mean utility. The fifth relates it to interpretability of test scores.

It would be difficult to state in words a core of meaning common to all the various definitions of test validity, of which the foregoing is only a sample. Such a conceptual definition, even if it could be formulated satisfactorily, would probably be too abstract to contribute significantly to more effective test validation. What the test developer needs is an operational definition.

Further, the generality of some of these definitions suggests that in the minds of their authors test validity is almost synonymous with test value. But if validity does mean value, then reliability, convenience in use, adequacy of norms, and even the availability of alternate forms become aspects of validity, and we are left without a term for what Gulliksen and Cureton mean by validity. Using the same term for a variety of concepts leads to serious semantic confusions and to procedural pitfalls as well.

TYPES OF VALIDITY

A closely related indication of difficulty with the concept of validity is seen in the diverse forms it must assume to fit different situations. The APA and the AERA Technical Recommendations mention four types of validity: content, predictive, concurrent, and construct. Two of these, content and construct, have little in common with the other two, or with each other. Anastasi discusses face validity, and factorial validity in addition to content validity, and various types of empirical validity. Gulliksen has discussed intrinsic validity, and Mosier analyzed face validity into validity by assumption, validity by definition, the appearance of validity, and validity by hypothesis.

Again it may be said truly that these types of validity have some common conceptual elements, but the differences are striking. To encompass all the diverse varieties of validity requires an extremely

loose and general definition of the basic idea of validity. It is easy to agree with Guilford that "The question of validity has many facets and it requires clear thinking not to be confused by them." Perhaps one could go farther and suggest that even clear thinking in the frame of reference of the present conceptual structure of validity may not lead to common understanding of a single concept, nor to effective operational use of it. Perhaps what we really need is not clearer thinking about validity, but rather a more concrete and realistic conception of the complex of qualities which make a test good.

EVIDENCE OF VALIDITY

A third indication that all is not well with validity is found in this strange paradox. While almost every test specialist agrees that validity is the most important quality of a mental test, almost all of them lament the general inadequacy of test validation. Nearly 30 years ago, in the early years of objective testing, Ruch made this comment:

There are in use today at least one thousand different educational and mental tests. Convincing critical and statistical data on the validity, reliability, and norms of these measures are available in probably less than 10 per cent of the cases.

One might reasonably expect that the situation would have improved in the intervening years, but this seems not to have happened. In a spaced sample of reviews of 20 tests in the *Fifth Mental Measurements Yearbook* only one was found in which the reviewer judged the evidence of validity to be adequate. Ten tests were criticized for lack of evidence of validity. Nine reviewers made no comment about the validation of the tests they reviewed. This in itself is surprising, if validity is indeed the most important quality of any mental test.

Some years ago, W. J. Cameron in one of his *Ford Sunday Evening Hour* commentaries observed that when someone tries to do a job the wrong way, nature often teaches him his error by refusing to let the job be done. Our failure to demonstrate consistently that our tests possess the quality we value above all others may mean that we have used the wrong approach in trying to gain evidence of it.

IS VALIDITY ESSENTIAL

A fourth suggestion that something may be wrong with the mental tester's concept of validity is that corresponding problems of validation seem to be almost nonexistent in the realm of physical measurements. Norman Campbell, P. W. Bridgman, and others have written extensively on the measurement of physical properties, but one searches in vain through their writings for a discussion of the validity of physical measurements. They show much concern for operational definitions of quantitative concepts, for limitations on the measurability of certain properties, and for accuracy of measurement. But the question of the *validity* of a measuring procedure seems to arise only incidentally and indirectly. For some properties, such as the hardness of solids or the viscosity of fluids, different methods of measurement yield inconsistent results. But modern physical scientists seem never to ask which of the methods of measurement is the more valid. One is moved to wonder why this difference between mental and physical measurement. Is it possible that we have fallen into a trap of our own devising when we find it so difficult to validate our mental tests? Have we, in Berkeley's words, "first raised a dust and then complained that we cannot see?"

Refinements in the measurement of distance—by the interferometer—or in the measurement of time—by the atom clock—are not justified on the basis of superior validity, that is, as closer approximations to the measurement of true distance or true time. They are regarded rather as improvements because they permit reproducible measurement to smaller fractions of existing units of measurement, which is to say that they are justified on the basis of superior reliability.

When a shortcut substitute for some more elaborate standard method of measurement is proposed, the question of the validity of the substitute method does arise with logical legitimacy. In such a situation the concept of validity is simple, and the meaning of the term is clear. We will argue for retaining this concept of validity and of restricting the term to this concept. But to ask about the validity of the basic method of measurement, which provides the operational definition of the thing being measured, would seem to most physical scientists as it does to us to be asking a meaningless question.

Why the Difficulties

SCIENTIFIC ADEQUACY

These observations suggest that the concept of validity itself may be weak scientifically. Most of the definitions of validity can be shown to be derived from the basic notion that validity is the degree to which a test measures what it is supposed to measure. But how does one know what a test is supposed to measure? On a superficial level, perhaps, it may be suggested by the test title—academic aptitude, mathematics achievement, or social studies background, for example—but these suggestions are by no means definitive.

Does the criterion tell us what the test is supposed to measure? It might if criteria were given to us. Usually they are not. They have to be *devised*, often after the test itself was constructed. Toops has said:

Possibly as much time should be spent in devising the criterion as in constructing and perfecting the test. This important part of a research seldom receives half the time or attention it requires or deserves. If the criterion is slighted the time spent on the tests is, by so much, largely wasted.

The ease with which test developers can be induced to accept as criterion measures quantitative data having the slightest appearance of relevance to the trait being measured is one of the scandals of psychometry. To borrow a figure of speech from Thorndike, they will use the loudness of the thunder as a criterion for their measurements of the voltage of the lightning. Even in those rare cases where criterion measures have been painstakingly devised, the validity of the test is not determined unless the validity of the criterion has been established. This requires a criterion for the other criterion, and so on *ad infinitum*. We can pursue such an infinite regress until we are weary without finding a self-sufficient foundation for a claim that the test is valid. It is an unhappy fact that the general conceptual definition of validity provides no firm basis for operational definitions of validity.

PHILOSOPHIC ADEQUACY

The concept of validity is also weak philosophically. It reflects a belief in the existence of quantifiable human characteristics, such as

intelligence or skill in arithmetic, independent of any operations used to measure it. Philosophers call this point of view realism but most of them now agree that it is not very realistic. One of Einstein's major contributions was to point out that the concept of time is scientifically meaningless until the clocks used to measure it have been described. As Henry F. Kaiser said in his review of a measurement book recently:

Chapter 2 repeatedly exhibits the philosophically naïve faith that there "exists" an "actual" or "true" scale for a particular phenomenon; the author seems to assume a degree of absolute truth inherent in nature which went out of style in the nineteenth century.

This naïve faith in the pre-existence of a quantity to be measured is basic to the general conception of validity.

You may recall the story of the three baseball umpires who were discussing their modes of operation and defending their integrity as umpires. "I call 'em as I see 'em," said the first. The second replied, "I call 'em as they are." The third said, "What I call 'em makes 'em what they are." In philosophical terms, the first was an empiricist; the second, a realist; and the third, a positivist. I should like to see test developers be less individualistic in their positivism than baseball umpires are at times, but I think they should be positivists rather than realists. Neither a strike in baseball nor scholastic aptitude in testing is a useful concept until it has been defined in operational terms.

Many of those concerned with mental measurements however, persist in being philosophical realists. They tend to endow abstractions with a real existence. They think of a real trait which "underlies" a test score, and which is meaningfully there even though their best efforts to measure it will never be more than approximations. They think of intelligence as really existing independent of any operational definition such as those provided by the Binet, the Kuhlman Anderson, or the Wechsler. They seek to use tests to discover what critical thinking or creativity really are instead of using the tests to define what they mean when they use such terms. They have not yet learned that realistic philosophy is productive mainly of verbal discourse, and that it must be shunned if mental measurement is to advance.

So long as what a test is supposed to measure is conceived to be an ideal quantity, unmeasurable directly and hence undefinable operationally, it is small wonder that we have trouble validating our tests. Only if we are willing to accept some actual test, or other actual method of obtaining criterion measures, as a basic (if somewhat arbitrary) operational definition of the thing we wish to measure, and only if we have some other test or measurement procedure that we wish to check against this standard, do we find the concept of test validity useful. Further, if the test we propose to use provides in itself the best available operational definition, the concept of validity does not apply. A basic definition needs to be clearly meaningful, but it does not need to be, and indeed it cannot be validated.

One of the by-products of the realistic philosophy is mistrust of appearances and a reverence for the concealed reality. What a test or test item really measures, we warn ourselves, may be quite different from what it appears to measure. But how a person can possibly determine what it really measures without observing something that it appears to measure is never clearly explained. Those who analyze batteries of tests to determine the "underlying" factors trust appearances of what a test is measuring very little, but even they fall back on appearances when they must name the factors discovered or provide verbal descriptions of them.

The source of our concern over the deceitfulness of appearances is probably that what a test appears to measure sometimes seems to be different to different observers or when viewed in a different light. If we resolve not to trust any appearances at all the problem vanishes, but so does our confidence in the test (and probably our sanity as well). A better course of action is to seek to understand why the appearances were not consistent, and to find an interpretation which makes them consistent.

Mistrust of appearances, in turn, leads one to seek completely empirical and deductive procedures of test validation. But completely empirical validation is seldom possible. Strictly speaking it is impossible in principle. We cannot escape judgment regarding the choice of a criterion, nor can we escape appearances (i.e., observations) in getting criterion data. To avoid an infinite regress of criterion validations one must stop somewhere and accept or proclaim an arbitrary definition of the thing to be measured. Unfortunately

this is seldom done. What happens more often is that we accept highly questionable criteria, obtain discouragingly low correlations, and finally give the whole thing up as a bad job.

OVERGENERALIZATION OF VALIDITY

A third possible explanation for difficulty with validity is that the concept is too broad. If it is made synonymous with value, or utility, or meaning, if it is made to apply to all mental tests including those used to describe persons or control educational processes as well as those used to predict future achievement, it must obviously have many different meanings. Now the trouble with using the same term to mean a variety of different things is that the meanings tend to get tangled up with each other. When the word is used in one particular sense, connotations appropriate to its use in other senses tend to hover about it and suggest irrelevant procedures.

In the case of the term validity, we tend always to expect evidence in the form of a validity coefficient even though such coefficients are completely appropriate only to tests used as convenient operational substitutes for more tedious, if somewhat more precise, standard measurement procedures. But when tests are used to describe educational achievement, or to assist in the control of the educational process, validity coefficient usually are quite irrelevant. The fact that they are not naturally relevant in these situations may account for some of the difficulty we encounter in trying to obtain data from which to calculate them. The obvious natural criteria we need simply do not exist in the real world, and must be conjured up from the realm of abstract ideals. Perhaps this is why evidence for the validity of educational tests is so often inadequate and unsatisfactory. Perhaps the notion of correlating test scores with criterion scores to obtain a basic index of test quality has been overgeneralized. Perhaps we have often sought to use it in situations where it does not logically apply.

It may even be that some of us, unconsciously perhaps, are glad to honor with our words a procedure of test validation which has limited applicability in practice. By so doing we exhibit our good intentions. If the procedure will not work in the absence of a good criterion, and if a good criterion is unavailable, we are excused from

further effort to demonstrate test quality. We also have, in the well recognized shortcomings of available criteria, a convenient scapegoat for the lack of good evidence of test quality. It may often be convenient to sweep the problem of test validation under the rug of inadequate or unavailable criteria, especially when we promise ourselves and others to work to get better criteria when we can find the time.

What Is a Criterion?

At this point it may be appropriate to ask what, after all, is the difference between test scores and criterion measures? Is the difference one of substance or only one of function? In the case of predictive validity the distinction is fairly clear. Test scores come first. Criterion measures are obtained later. In the case of concurrent validity the distinction gets blurred. One distinction suggested by frequent practice, is that criterion measures should be ratings based on direct observations of behavior under presumably natural conditions. This would serve to distinguish them from test scores, which are almost always based on assessments of output under carefully controlled and hence somewhat artificial conditions. But ratings based on direct observations of behavior have serious and well known psychometric shortcomings. This limits their value as criteria.

Indeed the limitation may be more serious than is commonly realized. Though it has often been done, it makes little sense to judge the accuracy with which a test does the job it is supposed to do by checking the scores it yields against those obtained from a less accurate measuring procedure. If a new method of measurement involves a better (and hence different) definition of the trait to be measured, it obviously makes no sense to judge its quality on the basis of degree of agreement with inferior measures. If the new method does not involve a better definition, but only more precise observations, it does make sense to require that the new agree with the old so far as their respective reliabilities will permit, but in this case it is hard to see the old, inferior measure as a standard or criterion for judging the quality of the new. If the criterion is used as a standard for judging the accuracy of the scores from the test, it

should always exemplify a measurement procedure clearly superior to (i.e., more relevant and precise than) that embodied in the test.

In theory this could provide a useful distinction between test scores and criterion measures. In practice it seldom does. What usually happens is that the test developer pours all the skill, all the energy, and all the time he has into the process of making an outstanding test. He has none left over to spend on obtaining measurements "clearly superior" to those his test will yield, and under the circumstances would have no stomach for the task anyway. Small wonder that many good tests go unvalidated or poorly validated by conventional psychometric standards.

PREDICTIVE VALIDITY

Predictive validity has long been recognized as one of the standard types, if not the standard type of validity. Cronbach, Mosier, and others have developed the idea that the purpose of all measurement is prediction. There is a special sense in which this is true, though the surveyor or the analytic chemist might be surprised to find himself in the same occupational class as the weather forecaster. Perhaps the statement "All measurement is for prediction" belongs in the same category as the statement "All education is guidance" or even "All flesh is grass." There is a degree of truth in such statements, but if they are taken too literally they can be seriously misleading. If the predictive function of measurement is regarded as the sole function, it leads to the highly questionable conclusion that the best way to judge the quality of a measurement of something is to determine how accurately it predicts something else.

Why should the quality of a Test X as a measure of Trait X be judged by how well it predicts Trait Y when Y is a function not only of X but also of Z, W, and possibly a host of other factors? Is it reasonable to judge the quality of a barometer solely, or even mainly, by the accuracy of the weather forecasts which are made with its help? Or to consider the matter in another way, is it reasonable to suppose that Test X should by itself be a good measure of Trait Y, when Test X consists of verbal analogies, arithmetic problems, etc., while Trait Y is ultimately measured by grades assigned by a variety of teachers in courses from Art to Zoology?

Scores on Test X may indeed be related to measures of Trait Y, and the size of the correlation may indicate, in part, how useful Test X is for a particular task of selection. But loose logic is involved if that correlation is used as a measure of the validity of Test X as a measure of Trait X. An academic aptitude test does not purport to *measure* academic success. It should not claim to do more than part of the job of *predicting* academic success.

SPECIFICITY OF VALIDITY

Validity, test theorists agree, is specific—specific to a given group of individuals tested, to the treatment given them, and to a given purpose for testing (or to a given criterion). Anyone who uses a published test is almost certain to give it to a different group than the one on which it was validated. For any user's group the test may be more or less valid than it was for the test author's tryout group. Quite possibly the user may even have a somewhat different purpose for testing than the test author had in mind. His criterion may be different. Again this means that the test may be more or less valid than the author reported. Under these conditions, how can a test author possibly publish fully adequate data on validity? The best he can do is to report validity under certain clearly specified and carefully restricted conditions of use. For the majority of possible uses of a test, validation becomes inevitably a responsibility of the test user. There is thus an element of unfairness in the common complaint that test publishers fail to provide adequate data on validity.

Alternatives

MEANINGFULNESS AND VALIDITY

Whether or not you are prepared to agree that validity has serious shortcomings as the primary basis for judging test quality, you may now be interested in what alternatives might be proposed to replace it. What basis for judging test quality would be better than validity? Cronbach's definition cited earlier, may provide a clue. He said: "The more fully and confidently a test can be interpreted, the greater its validity." The interpretability of a test score depends on its meaningfulness. We would suggest that meaningfulness replace

validity in the usual lists of major desirable characteristics of a measuring instrument. Before this suggestion is laughed out of hearing, consider what it implies.

One, but only one, of the kinds of information that help to make test score meaningful is the relation of those scores to other measures of the same persons. When tests are used to predict, or when they are used as convenient substitutes for more exact but more laborious measurement procedures, validity coefficients expressing the relation between test scores and criterion measures may be the most essential basis for meaning. Hence we are not proposing that either the term or the concept of validity be abolished but only that they be restricted to situations in which independent criterion measures are feasible and necessary.

Relationships of test scores to other measures can also add meaning to the test scores even when the other measures do not constitute legitimate criteria. When a test is used in a battery, knowledge of intercorrelations among the scores adds to the meaningfulness of the scores from each test. Such intercorrelations show how much the various tests measure in common, and how much independent information they provide. Campbell and Fiske have suggested a special technique for using this kind of information. From a "multi-trait-multimethod matrix" of intercorrelations they secure data on which to base "convergent and discriminant" test validation. Construct validation also depends on relations between measures of various kinds, but thus far it has been of more direct interest and value to the psychological theorist than to the psychometrist.

Unless a measure is related to other measures it is scientifically and operationally sterile. The validity fallacy arises from the assumption that the relation of the measure to one single other measure (the criterion) is all important. The concept of construct validity has helped to break down this unfortunate stereotype.

Operational Definitions. What of the other kinds of information that help make test scores meaningful? Most important of all, scientifically, is a description of the operations used to obtain the scores. Operational definitions have always been basic to the meaning of measurements of length, mass, time, and other physical quantities. Such operational definitions should be basic to mental measurements as well. They would be, I am persuaded, had we not been misled by an overgeneralized concept of predictive validity.

Operational definitions of some kinds of test scores, such as speed scores in typewriting, ability scores in spelling, or vocabulary knowledge scores, are not particularly difficult to formulate. For other test scores, the problems seem more formidable. We must acknowledge that the excellence of many current tests has resulted more from the skilled intuitions of the test constructor than from preconceived excellence of design, recorded in truly controlling test specifications. But there is no apparent reason why an adequate operational definition of the score from any test should be impossible. Such a definition obviously must cover the critical procedures in test construction, in test administration, and in scoring. The development, use, and publication of such operational definitions would, I am persuaded, not only make the test scores more meaningful, but would lead us rapidly to the production of better tests.

Reliability and Norms. There are two other types of information which contribute substantially to the meaningfulness of test scores. These have to do with the reliability of the scores and with the norms of performance for representative groups of examinees. A completely unreliable score is completely meaningless. A perfectly reliable test score is almost certainly meaningful, though it may not be particularly significant or useful.

The importance of norms in making test scores meaningful requires no defense here. In the case of most educational tests they are highly useful. In a few special cases they may be unimportant or even irrelevant.

IMPORTANCE AND CONVENIENCE

The stress we have placed on meaningfulness of test scores, subordinating validity, reliability, and norms to it, does not mean that it can be regarded as the sole basis for judging the quality of a mental test. There are two other very important elements. One is the importance (usefulness as a basis for effective or satisfying behavior) of the knowledge or abilities required by the test. The other is the convenience of the test in use. The many factors which contribute to the convenience of a test have been well outlined by numerous authors.

A measurement can be completely meaningful and still be completely useless. For example, the number of hairs on a person's head

is an operationally definable measurement. It can be related to other measurements of a person such as his age or his IQ. We could estimate its reliability and get norms for it. But it would remain, so far as I know, an almost useless measurement, one of little or no importance. Quite properly I think, critics of current educational tests are as much concerned with the importance of what the test is measuring as they are with the meaningfulness of the scores or with the convenience of the test in use.

Conclusions

It may be helpful now to summarize in outline form the characteristics which we regard as determining the quality of a mental test or measurement procedure. They are:

1. The importance of the inferences that can be made from the test scores
2. The meaningfulness of the test scores, based on
 - a. An operational definition of the measurement procedure
 - b. A knowledge of the relationships of the scores to other measures, from
 - i. Validity coefficients, predictive and concurrent
 - ii. Other correlation coefficients or measures of relationship
 - c. A good estimate of the reliability of the scores
 - d. Appropriate norms of examinee performance
3. The convenience of the test in use

Must all tests be valid? If the term "valid" is not to be made synonymous with the term "good," if validity is a clearly defined concept which can be quantified by finding the correlation between test scores and criterion measures, then the answer is clearly "no," on the basis of the considerations discussed in this paper.

These views may be wrong. If so, and if the current conception of validity is philosophically sound and operationally useful, let us in the name of intellectual honesty, support this claim with some good solid evidence. The time is long past for lame apologies and prolix rationalization of failure to demonstrate that good tests have the quality we have said is more important than any other. Perhaps we should recognize the age-old alternatives so far as validity is con-

cerned. Either put up the evidence or withdraw the claim. It is my view that in general, we have not and will not be able to put up satisfactory evidence. On the other hand we should not stop being concerned about test quality. What is proposed here is that we stop beating our heads against a stone wall and step back to look for a way over it or around it. There is one, I think, and this article has attempted to elucidate it.

Having followed the argument thus far some will say: "You still want valid tests. All you have done is to propose a different term, meaningfulness, to replace validity." This is surely not what I have been trying to do. I hope that your time has not been wasted in reading one side of a purely lexical debate. I hope that these efforts may contribute to the adoption of a more appropriate and productive procedure than validation has been for determining the quality of a test.

References

- AMERICAN EDUCATIONAL RESEARCH ASSOCIATION, Committee on Test Standards. *Technical recommendations for achievement tests*. Washington, D.C.: AERA, 1955.
- AMERICAN PSYCHOLOGICAL ASSOCIATION, *Technical recommendations for psychological tests and diagnostic techniques*. Washington, D.C.: APA, 1954.
- ANASTASI, ANNE, *Psychological testing*. New York: Macmillan, 1954.
- BRIDGMAN, P. W., *The logic of modern physics*. New York: Macmillan, 1927.
- BUROS, O., (Ed.), *The fifth mental measurements yearbook*. Highland Park, N.J.: Gryphon, 1959.
- CAMPBELL, D. T., & FISKE, D. W., Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol. Bull.*, 1959, 56, 81-105.
- CAMPBELL, N. R., *Foundations of science*. (Originally published as *Physics: The elements*.) New York: Dover, 1957.
- CRONBACH, L. J., *Essentials of psychological testing*. (2nd ed.) New York: Harper, 1960. (a)
- CRONBACH, L. J., Validity. In C. W. Harris (Ed.), *Encyclopedia of educational research*. New York: Macmillan, 1960. (b)
- CURETON, E. E., Validity. In E. F. Lindquist (Ed.), *Educational measurement*. Washington, D.C.: American Council on Education, 1951. Pp. 621-694.
- EDGERTON, H. A., The place of measuring instruments in guidance. In Wilma T. Donahue, C. H. Coombs, and R. M. W. Travers (Eds.), *The*

- measurement of student adjustment and achievement. Ann Arbor, Mich.: Univer. Michigan Press, 1949. Pp. 51-58.
- GUILFORD, J. P., *Fundamental statistics in psychology and education*. (3rd ed.) New York: McGraw-Hill, 1956.
- GULLIKSEN, H., Intrinsic validity. *Amer. Psychologist*, 1950, 5, 511-517. (a)
- GULLIKSEN, H., *Theory of mental tests*. New York: Wiley, 1950. (b)
- KAISER, H. F., Review of Virginia L. Senders, *Measurement and statistics*. *Psychometrika*, 1960, 25, 411-413.
- LINDQUIST, E. F., *A first course in statistics*. (Rev. ed.) Boston: Houghton Mifflin, 1942.
- MOSIER, C. I., A critical examination of the concepts of face validity. *Educ. psychol. Measmt.*, 1947, 7, 191-205.
- MOSIER, C. I., Batteries and profiles. In E. F. Lindquist (Ed.), *Educational measurement*. Washington, D.C.: American Council on Education, 1951.
- RUCH, G. M., Recent developments in statistical procedures. *Rev. educ. Res.*, 1933, 3, 39-40.
- THORNDIKE, E. L., Measurement in education. In, *Twenty-first yearbook of the National Society for the Study of Education*. Bloomington, Ill.: Public School Publishing Company, 1922.
- TOOPS, H. A., The Criterion. *Educ. psychol. Measmt.*, 1944, 4, 271-297.

WILLIAM M. LITTELL
San Francisco State College

Further Comments on Test Validity*

Many textbooks avoid controversy. However, if the students can grasp the basic issues involved, it may be well to let them know that educational and psychological problems are more complex than textbook presentations may indicate, and that even the experts do not always agree. The "proper" use of standardized tests (for example, for intelligence, aptitudes, achievement, personality) has always been an issue. When the

* Published with the permission of the author.

objective testing movement reached its zenith of popularity before World War II, students were routinely subjected to batteries of tests. The school personnel who administered, scored, and interpreted the tests frequently knew little about their construction and inherent limitations. Test scores became sources of divinely revealed truth. The IQ tests particularly were enshrined, and a student's score became one of his indelible identifying characteristics. With the current national talent hunt there has been a resurgence of testing in the schools, and a student's educational advancement or college admission often hangs on the scores that he receives on standardized achievement and ability tests. The ease with which tests can be administered and scored, and the superficiality with which they can be interpreted, have made them popular rituals in the schools. Somehow the test becomes more important than the student who takes it and the test results more valued than the empirical realities of Johnny's behavior.

Littell's criticism of Ebel's view of validity and the use of tests is that it may perpetuate this state of affairs. Littell does not want the test-taker subordinated to the test. Nor does he want the psychological interpretation of the student's behavior to rest solely on the results of the test. The choice of appropriate test batteries must be tailored to individual need and rest on professional judgment. Even then the results are not revealed truth; but they do furnish the psychologist with some hypotheses about the nature and causes of an individual's behavior, and these hypotheses must be checked in a specific set of circumstances. In this latter sense, test validity (or the particular usefulness of a test in given circumstances) remains an important concept.

In considering the matter of validity, Littell has expressed a view about the present state of psychological knowledge which may differ from what Ebel has at least tacitly assumed. At present, we have much more theory in psychology than we have experimental tests of theory. Psychological theory often gives us "images" of behavior to which we become subjectively attached. When this occurs, we no longer trust the feedback that we can obtain when we try to apply the theory to particular

individuals and situations. The rationale upon which tests are built is often such a body of untested theory or assumptions. To use the tests without checking them against the observable facts of Johnny's behavior is to have purely subjective faith in tests and theory. In the present state of our science, psychological tests often lack the valid theoretical underpinnings assumed in their indiscriminate use. Until we have much more specific information about how well the tests do their jobs under particular conditions, little trust can be placed in test scores. It should be clear, however, that Littell is not demeaning psychological theory and tests; he is simply urging a more realistic view of their present limitations.

It might be of some interest for the student to organize two testing programs: one should follow the suggestions of Ebel and the other the suggestions of Littell. How would you "test" the programs to discover which was more effective? Also, how does Littell's view of testing technology or practice agree with Melton's view about the present state of educational technology (p. 21)?

The concept of test validity has been subjected to increasingly close scrutiny over the past few years (e.g., A.P.A., 1954; Cronbach, 1960; Littell, 1960). This scrutiny has reflected and indeed is an outgrowth of the growing understanding of the process of theory development, use and evaluation within science in general. This particular discussion has been stimulated by Ebel's recent articles (1961) which questions very seriously the concept of validity as it is currently employed by test developers and users. Certainly, as Ebel points out, the concept of validity as it is currently applied and discussed, is confused and contradictory and based on an outmoded philosophy of naïve realism; it is of great concern to test developers, but difficult to apply and often receives only lip service. If the case were closed at this point, the test user would be left in a real and very serious "validity dilemma."

Ebel proposes a way out of the dilemma through the concept of "meaningfulness" which he suggests should replace validity as the

primary factor to consider in evaluating a test. Meaningfulness, for Ebel, arises from knowledge of: (a) an operational definition of the measurement procedure; (b) the relationship of the scores to other measures (validity coefficients, predictive and concurrent, and other correlation coefficients or measures of relationship); (c) a good estimate of the reliability of the scores; (d) appropriate norms of examinee performance. This factor of meaningfulness, along with convenience in use and, a third criterion, "the importance of the inferences that can be made from the test scores (p. 646)" constitute the basic criteria offered by Ebel upon which one should base his choice and use of a psychological test. Five and one third pages are devoted by Ebel to the discussion of the limitations of the current concepts of validity, a little more than one page to the presentation and discussion of "meaningfulness" and only two short sentences to the mention of what the writer sees as the critical issue of test validity, *the inference from test scores*. In this sense, Ebel has really not taken up the issue of *validity* at all. With a casual reading, one might be left with the impression that somehow by knowing how a test was developed and administered and a variety of correlation coefficients with other measures (usually other tests), the process of inference is clear and unambiguous, and needs no further thought or evaluation.

The following discussion takes up the issue of inference and compares some of the implications which emerge with some of the points Ebel has made. In general, this discussion takes a position which reflects several basic ideas about psychological test validity:

1. Test validity is best understood and the issues are most clear when considered within the context of the choice and use of psychological tests by those actively engaged in using them.

2. All matters of test construction (convenience, reliability, norms, "meaningfulness," etc.) are important only in so far as they bear on the process by which the test user formulates, applies, and evaluates the inferences he makes from the test scores.

3. In the last analysis the value of a test depends upon the degree to which it aids the user in reducing his error in anticipating other behavior more relevant to his professional activities.

4. While this article is written with the standard psychological test in mind, the issues brought out and the positions taken are

equally applicable, although perhaps with different emphasis, to any situation in which one chooses to measure or identify a variable in a conceptual system in a particular way.

The Use of the Psychological Test

Whether the user of a psychological test is operating within a clinical, educational, industrial, or research setting, he engages the use of a psychological test for specific reasons highly related to his professional activities. Aside from those thousands of tests administered merely to fill in blanks on forms which have become functionally autonomous, the user employs a given test because he wants to know more about the person than simply his score. He wishes to be able to infer, to generalize, to gain information about a person that would not be available as easily, or cheaply, or efficiently, or quickly had he not administered the test.

In the usual situation, the professional psychologist or educator is faced with the necessity of formulating and selecting from among possible plans of action through which he hopes to achieve his therapeutic, remedial, educational, or advising goals. His effectiveness as a professional rests very heavily on his ability to formulate and select the "best" plan. The plans he makes and the bases for his choice among them depend in part on such practical considerations as the problem presented, his professional purpose, the facilities available, ethics, etc. The factors having to do with the specific people involved, however, are the ones which rightly demand his deepest consideration; he must know as accurately as possible how *the people themselves will respond to them*, and as an aid to this end he employs the psychological test. He chooses to give a test rather than to gather information in some other manner, and to give this particular test (or tests) because he has reason to believe that in this way he can maximize his efficiency in gathering information that will be relevant and helpful in formulating and choosing from among these alternative courses of action open to him.

This approach has interesting implications. One should give a test only if he cannot select with a *sufficient degree of certainty* from among the alternative courses of action on the basis of information already at hand; if nothing the test could show would alter his

choice, then giving the test is a waste of time and money. Tests should be given for a specific purpose growing out of the user's need in *that particular instance* for specific information. (The valid "exploratory" use of tests must be recognized, of course, when indeed one does not have enough information to begin even to formulate alternative courses of action.)

The problem remains of how a test user is to assess the ability of a particular test to do a specific job in any given situation. At present this is at best a very complex and often unrewarding task.

The Basic Need for Empirical Validation

Apparently central to Ebel's approach is the suggestion that there is some way *other* than through empirical validation by which a test may be established as truly useful. Responsible use of a test *must*, in the last analysis, be tied to empirical observation. Somewhere along the line in his use of a test the user must make some statements that can be checked by direct observation. The final question as to the validity of the test as used is answered only through these observations. If there is no way to tell whether one is right or wrong (even though this evidence may be quite distant in time and connected only by inference) then one might as well let his fantasy run unchecked and answer only to his aesthetic and/or libidinal impulses in his choice and use of a test. (Better still, he can fall back on common usage: "I may not be right, but no one will dare tell me I'm wrong.") Ethically, however, he is committed to make it clear to others what he is doing.

For the test user the problem of test validity is relatively simple when he has a clear idea of what he wants to predict and has available sound empirical evidence that bears directly on the ability of the test to do in fact what he wants it to do under his particular set of circumstances. Very seldom is this the case, however, and usually the test user is forced (or even prefers) to infer beyond any clearly applicable predictive studies. He does this usually by finding a test that will (apparently) measure some supposed enduring characteristic of the individual which is not directly observable and for which no one observable criterion exists, but which his theory (or "understanding") tells him is important (i.e., related to the behavior

he ultimately wishes to predict). We assume, of course, that the user is interested in something more than achieving subjective "closure," something more than filling in the gaps in his own subjective understanding of his client, subject, or student. We suspect that this feeling of subjective understanding *can* come about, for instance, simply by fitting the phenomenon into an old, familiar context which may, if judged on subjective criteria alone, be not only objectively misleading but quite wrong.

The question of the ability of the test used in this manner to do what the user wants it to do must be dealt with in terms of some level of construct validation. At present this is to open a Pandora's box of problems, for under these circumstances the user's faith in the inferences made from the test must rest upon his assessment of the "validity" of *both* the test as a measure of the variable and of the *whole conceptual system* in which the variable takes on the meaning from which future predictions of behavior are made. This puts the psychological test user in a very dangerous situation, for at present under the best of circumstances he is forced to operate with a conceptual system in which clinical, personal, educational, and cultural folk lore combine in almost unknown quantities with the little objective evidence available.

At some future date the term "validity" might be used to refer to the confidence with which one can substitute the score of a particular psychological test for a term in a pre-existing, well substantiated, internally consistent theory or set of laws. At present, however, psychology, especially in the applied fields, has only the barest beginning of such theories of sets of laws. Strictly speaking, under current conditions, it is meaningless to look for the construct validity of a test measure except in the most gross terms. The situation is difficult enough when we speak of the construct validity of a measure of intelligence (e.g., Littell, 1960); consider the problems involved in assessing the construct validity of a measure of such concepts as *social presence* or *extraversion*.

For the test user, this means that he cannot look upon his theory as more than a guide, a source of hypothesis to be tested. He must approach these measures of an hypothetical construct with extreme caution, and continue to investigate critically the "validity" (internal consistency, agreement with facts, etc.) of the conceptual

system(s) he uses in his professional endeavor. The long term goal for psychology as a science is to develop theoretical structures of sufficient definiteness, etc., that their "validity" and the validity of the measures of the constructs they involve can be investigated. The present goal for the users of psychological tests must be to compensate for this lack of easily generalized validity by sticking as closely as possible to short term, unglobal, "checkable" uses of tests.

The Problem of the Criterion

Many people appear to share Ebel's concern over the failure to find "adequate" criterion measures by which to establish a test's validity. At this stage of theory development we would be surprised if in fact the situation were other than this. Whenever one wishes to measure a variable that has meaning beyond that which can be contained in any one observation (and this is the usual case), one must deal with the process of construct validation. While this is especially true with "psychological" variables, it is true even with such supposedly obvious terms as *arithmetic ability*, *immediate recall*, *achievement in social studies*, etc. The search for an "ultimate" criterion is bound to fail.

When a criterion against which a test is to be validated is chosen, of course its reliability, relevance, etc. must be investigated, and of course at this stage of development of psychological theory and measurement the criterion will be found to be just as lacking as any test in the degree to which it is isomorphic with the rich, involved (and usually somewhat contradictory and ambiguous) cognitive structure behind the behavior of the investigator. But one cannot throw in the sponge and selectively *avoid* this disappointing but *essential* empirical anchoring of any scientific theoretical structure.

A mistrust of observation results primarily when one expects too much from observation; the observation of the relation between the test and one criterion event can establish the final "validity" of the test only when one is interested in predicting *that* particular event. As long as test users continue to operate on the basis of "folk lore" theory (be it personal, clinical, or educational) no appeal to data will really capture the essence and full meaning of the concept sought.

The Essential Mistrust of Appearances

Ebel chooses to reject the suggestion of the disillusioned naïve realist that we mistrust appearances. It is the writer's opinion that the only way to make healthy use of a test is to have a firm and enduring mistrust of appearances. The validation of a test is not the final discovery of what the test (really) measures, no matter how it is phrased (e.g., the abilities it taps, the personality factors it reflects, the skills it measures, etc.), but is more the accumulation of enough empirical information about the test (e.g., factors which influence the scores, necessary conditions to be met in administration, the relationships into which it enters with other *behavioral* measures, etc.) to have reason to suspect that it will in fact do what we want it to do. Its use is to be based on this information, and not on the label it is given, or the classification into which it is placed by its appearances, whether these appearances are superficial and naïve evaluations of its content or detailed and complex descriptions of the development of the test.

There need be nothing "mysterious" about the use of a psychological test. In a very real sense the psychological test should be a tool for the professional psychologist or educator much like a hammer is a tool for the carpenter. One does not speak of the "validity" of a hammer; as one learns the trade, one finds out what can be done with it and how to use it in order to obtain the best results. Only for the novice must one label a hammer "an instrument for pounding." In other words, the test user also should base his use of a test on what he knows from experience (his own or that of others) the test can do.

The use of a test is in a very basic sense arbitrary; that is, unrelated to its name, its "classification," or any associations the psychologist might have to its content, development, etc. The same basic questions of validity must be dealt with no matter how the test is used. For example, an intelligence test could be used "validly" as a clock; a very rough and inefficient clock to be sure, but a clock none the less. We administer the intelligence test to a person at time t and then readminister it at some time Δt later. By reference to appropriate data (diminution of practice effects over time) a rough estimate of Δt could be inferred from the difference between the

two scores. Certainly it would be rough and inaccurate and very inefficient as a measure of time, especially when compared with other more standard time measures. But it could be expected to render an estimate more accurate than would be obtained by merely guessing at random from all possible times. As a measure of time the "difference between two intelligence test scores" should be rejected because it is less efficient and reliable and less useful than other measures; not because "an intelligence test does not measure time."

The Operational Definition of a Test

Ebel suggests that the operational definition of a test ("critical procedures in test construction, in test administration, and in scoring") be included as an important aspect of what he terms test "meaningfulness." There is no question as to whether this information should be available to any user of the test. There is a question left unanswered by Ebel's discussion, however, as to how the psychologist should make use of this information. An operational definition in and of itself, aside from aiding competent administration of the test, means very little. Once a measure is devised one need not know how or by what method unless there is reason to suspect that such information will be of help in the use of the test. The test could be distilled from the residue at the bottom of a witch's caldron if it were convenient and inexpensive to use, reliable and entered into a large number of significant relationships with other behavioral variables important to the psychologist. We suspect, of course, that there are some ways of devising a test to do a certain task that will have a greater chance of being fruitful than others, and many texts have been written about these suspicions. Once devised, however, it must stand only on the evidence which shows that it can *in fact* do what it was devised to do.

For the test user, an operational definition is perhaps best understood as a check point, an attempt by the developer of the test to spell out in as much detail as possible all of the factors which might bear on the test score. There are no hard and fast rules as to what should be included although, of course, there is considerable agreement. What is included in the operational definition of any test is

left up to the judgment of the test constructor as to whether from his knowledge, background, etc. any particular factor should be considered important to include. Few test constructors would bother to include a minute description of the colors of the shirts of the subjects upon which the test was developed, simply because there is no reason to suspect that this variable could have any effect on the test score. If he felt it were important and, better still, had evidence, then shirt color would become a legitimate part of the operational definition of the test.

Test Reliability and Standardization as Possible "Distractors"

Considerable effort and skill has gone into the development of techniques for increasing the reliability of psychological measuring devices. This is all well and good. Care must be taken, however, that due to the effort expended on it, reliability does not assume a position in the evaluation of a test out of keeping with its actual contribution to the user's faith in the test. A test can be used only to the degree that it is reliable. A sufficient degree of reliability is an *essential but completely insufficient* condition for the use of a test. It is a condition which must prevail *before* one can begin on the basic task of finding what in fact the test can do.

Tests are highly standardized and systematized methods of observation, and as such there is a great potential for strength and efficiency in their use. This usefulness must be built into the test and assessed empirically, however, for these tests also provide highly systematized patterns of behavior for the user, and therefore apparently may come to achieve the status of the compulsive ritual, an act which, it is well known, tends to be perseverated in spite of any and all feedback. In fact, it is just in such situations in which feedback is ambiguous and anxiety is high that we expect these magical rituals to tend to develop.

A Final Comment

There is in fact a "validity dilemma" in common psychological test usage today. When the test user departs from directly applicable empirical data he is on tenuous ground; conclusions from available

studies are overgeneralized and the psychological theory in use is all too often little more than a rough guide. Perhaps the best course of action for the test user is to recognize the essentially limited contribution of psychological tests to professional endeavor and to use them realistically as they are. One can face this fact and still find sufficient reason to involve tests as still another source of information (with many unique and helpful features) *the usefulness of which must be evaluated under any specific set of circumstances.*

In this light perhaps the best thing a test constructor or test company can do is to provide a convenient, inexpensive and highly reliable measure of behavior which evokes a rich set of associations (inferences) in the professional test user and which therefore provides a large number of hypotheses regarding the possible use of the test to be then checked in any specific set of circumstances. The test user has only one course of action open to him: to be continually aware that he is working with hypotheses to be tested. At no time can he stop and trust appearances, subjective certainty or current theory.

References

- AMERICAN PSYCHOLOGICAL ASSOCIATION, *Technical recommendations for psychological tests and diagnostic techniques*. Washington, D.C.: APA, 1954.
- CRONBACH, L. J., Validity. In C. W. Harris (ed.), *Encyclopedia of educational research*. New York: Macmillan, 1960.
- EBEL, R. L., Must all tests be valid? *Amer. Psychologist*, 1961, **16**, 640-647.
- LITTELL, W. M., The Wechsler Intelligence Scale for Children: review of a decade of research. *Psychol. Bull.*, 1960, **57**, 132-156.

[INDEX]

- Adkins, Dorothy C., 129, 567, 568,
575-586, 587
- Adult intelligence, temporal changes
in, 472-474
- Aids in problem solving, 234-235
- Aikman, L., 229, 235
- Air Force, U.S., 27, 33, 437, 534
- Allen, L., 459
- American Psychological Association, 7,
25
- Analysis of classroom tests, procedures
for, 586-613; for discriminating
power, 594-595; for relevance, 588-
594; of score discrimination, 595-596
- Anderson, S. B., 230, 231, 236, 237
- Appearances, mistrust of, 622
- Aristotle, 420, 422, 423
- Asch, M. J., 555-556
- Aspiration, levels of, 38-39
- Association, nature of, in learning
efficiency, 41, 42-44; response simi-
larity, 42-44; stimulus similarity, 42
- Associative factors of learning condi-
tions, 31, 34-35, 41-50; association,
nature of, 41, 42-44; inter-trial, 42,
47-50; intra-trial, 41, 44-47
- Attitude change, grades as reinforcers
in production of, 129-135
- Attlee, Clement, 444
- Audio-visual aids, 165, 381-415
- Austin, G. A., 215, 247
- Ausubel, David P., 9, 67-80, 86, 278,
279, 483
- Azrin, Nathan H., 182
- Backer, R., 104
- Barker, R. G., 118
- Battig, W. F., 232
- Bayley, Nancy, 456-477, 483
- Behavior change, prediction of, im-
proved, decision-making and, 55-58;
inferential, in preschool children,
271-277; theory, 23
- Bendig, A. W., 240
- Benedetti, D. T., 221, 231
- Bergum, 246, 247
- Berelson, 354
- Berkeley, 602
- Bernstein, B. B., 218, 235
- Binet, Alfred, 424, 426-427
- Birch, H. C., 92
- Blazek, N. C., 103
- Bloom, Benjamin S., 108, 210, 220, 239,
241, 568, 569-575
- Bloomster, 348-349
- Bogartz, W., 289, 290, 294, 295, 302
- Bolles, Robert C., 6, 9, 30-51, 52, 68,
85, 139, 183, 190, 211, 311, 383, 384
- Bostrom, Robert N., 84, 108, 129-135
- Breger, L., 289, 290, 294, 295
- Brenneise, S. H., 229, 231, 232
- Bridgman, P. W., 602
- Briggs, L. J., 320
- Broder, L. J., 220, 239, 241
- Brooks, L. O., 217, 221, 227, 234, 302
- Brown, J. S., 104
- Brown, Roger, 326, 338-355
- Bruner, Jerome S., 7, 210, 211, 213,
215, 247, 254-270, 278, 357, 436, 483,
512, 527, 551
- Brush, F. R., 233, 247
- Bryan, William Lowe, 330, 331
- Bugelski, R. B., 83-84, 107, 108, 118,
164, 382, 483
- Burack, B., 217, 234
- Burt, Sir Cyril, 165, 417, 419-435, 483,
512
- Buswell, G. T., 218, 239, 583
- Cameron, W. J., 601
- Campbell, Norman, 602, 610
- Carpenter, Edmund, 399
- Cattell, J. M., 339, 341, 345, 428
- Charles, D. C., 476
- Child, Irvin L., 84, 117-128, 157, 164,
483

- Children, curriculum planning, voice in, 75-77; development, educational practice and, 68-70; emotion of, educating, 10-20; intelligence of, growth of, 456-477; motivation in, manipulatory, 98-106; preschool, inferential behavior in, 271-277
- Chomsky, N., 267
- Chown, S. M., 213, 221, 224, 241, 242, 245
- Christal, R. E., 443
- Churchill, Winston, 444
- Cicero, 420, 422
- Claparède, 424
- Classroom, laboratory and, link between, 5-80
- Classroom tests, analysis of, procedures for, 586-613, for discriminating power, 594-595, for relevance, 588-594, of score discrimination, 595-596; operational definition of, 623-624; reliability and standardization of, as "distractors," 624; validity of, 597-613, alternatives, 609-612, comments on, further, 614-625, criterion, definition of, 607-609, definitions of, 599-600, essentialness of, 602, evidence of, 601, importance and convenience, 611-612, meaningfulness and, 609-611, overgeneralization of, 606-607, philosophic adequacy, 603-606, predictive, 608-609, reasons for difficulties, 603-604, scientific adequacy, 603, specificity of, 609, types of, 600-601
- Cobb, H. V., 229, 231, 232
- Code as communication, 327-330
- Cofer, C. N., 214, 225, 231, 243, 296
- Cognitive abilities of intellect, 440-443
- Cohen, Arthur R., 529, 530, 559-565
- Coladarci, A. P., 57, 65
- Communication, 323-380; code as, 327-330; improving, 335-338; language learning, 332-334; learning context, 330-332; logic, thinking, and teaching, 356-366; meaningful material, verbal context, and recall of, 367-380; psychologist's view of, 325-338; reading, a dispute about, 338-355; social class and modes of, 511-526, classifications and classified relations, 518-519, differences between, 514, framework and stylistic devices, organizing, 519-523, imagery between speaker and listener, correspondence of, 516-518, lower class, 514-515, 516-519, 520-522, middle class, 515, 517, 519, 520-523, perspective or centering, 514-515; understanding, 334-335
- Comrey, A. L., 236, 237
- Concreteness in problem solving, 228-229
- Conservation of memory, 269-270
- Convergent-production abilities of intellect, 447-449
- Conversation, content of, control of, 140-156
- Cooper, Devra, 182
- Corey, S. M., 62
- Corman, B. R., 217, 233
- Craig, R. C., 218, 279
- Crannell, C. W., 216, 217, 234
- Criterion, definition of, 607-609; problem of, 621
- Cronbach, L. J., 599, 608, 609, 616
- Crowder, Norman A., 137, 183-189, 200, 357
- Crozier, W. J., 94
- Cumulative constructionism, 260
- Cureton, E. E., 599, 600
- Curiosity, motivation, 86-97
- Curiosity-investigation motive, 95-96
- Curriculum, breadth of, 73-75; development, educational, objectives and, 569-575; planning, child's voice in, 75-77
- Darley, 533
- Darwin, Charles, 81, 423
- Decision-making, improved prediction of behavior change and, 55-58; relative validity of generalizations and, 54; specific, applying generalizations in making, 56-58
- Deese, 7
- Dembo, Tamara, 118
- Development, child, educational practice and, 68-70; cognitive, organization and, 77-80, human growth and, 67-80
- Dewey, John, 164, 210-211, 212, 255, 312, 357, 573
- Dickson, W. J., 114
- Differences, individual, 479-526; in growth rates, 468-472; intellectual ability, achieving and underachieving boys of high, 501-511; learning and, 23; psychology of, education and, 482-500; social class and modes of communication, 511-526
- Difficulty in problem solving, 233-234

- Discovery, act of, 254-270, intellectual potency, 259-263, learning the heuristics of, 267-269, memory, conservation of, 269-270, motives, intrinsic and extrinsic, 263-267; directed, learning by, motivating effect of, 277-288
- Discrimination, analysis of classroom tests for, 594-595; score, analysis of, 595-596
- Distributed practice, factor in learning efficiency, 47-48
- Divergent-thinking abilities of intellect, 444-447
- Dodge, 342
- Dolch, 348-349
- Doob, L. W., 130
- Douvan, E., 99
- Draguns, J., 246
- Dressel, Paul L., 585
- Drevidahl, J. E., 307
- Drive-reduction, 81-82; theory, 88, 91
- Drives, 82, 83-97
- Duncan, Carl P., 107, 209, 210, 211, 212-248, 271, 277, 311, 315, 319, 320
- Dunker, K., 224, 234
- Durkin, K., 99
- Ebbinghaus, Hermann, 311, 579
- Ebel, Robert L., 568-569, 570, 586-613, 615, 616-617, 619, 621, 623
- Eckerson, 343
- Edgerton, H. A., 599
- Edgren, R. D., 216
- Education, audio-visual, stimulus-response psychology and, 384-393; emotion of children, 10-20; implication of structure of intellect for, 454-456; objective of, and curriculum development, 569-575; psychology and, 1-3; psychology of individual differences and 482-500; TV research in decade ahead, 394-403
- Educational methods, technology of, learning and, 21-30; practice, *see* Practice, educational; process, measurement in relation to, 575-580
- Edwards, A. L., 315
- Einstein, Albert, 604
- Eisman, E., 217, 221, 227, 234
- Empirical validation of tests, need for, 619-621
- Engstrand, R., 227
- Episodic empiricism, 260, 261-262
- Erdmann, 342
- Estes, 386
- Evaluation, educational practice, procedures for, 58-67
- Evaluative abilities of intellect, 449-451
- Exploratory drive, 83
- Expository mode, 258
- External stimulation, 88
- Fattu, N. A., 220, 231, 246
- Faust, W. L., 235, 236, 240
- Faw, V., 554-555
- Ferster, Charles B., 182, 189-199
- Films, 381-393
- Fiske, D. W., 610
- Fixedness, functional, 224-227
- Flavell, J. H., 246
- Flesch, Rudolf, 343
- Flory, C. D., 475
- Forgus, R. H., 216, 217, 233
- Frankel, Edward, 481, 482, 501-511
- Freeman, F. N., 475
- French, R. S., 219, 233, 247
- Freud, Sigmund, 81
- Frick, J. W., 222, 245
- Frost, Robert, 400
- Frustration, 84, 95; effects of, personality and, 120-123; introspection variable, 123-125; kind of task, effect of, 125-127; quality of performance and, 117-128
- Gagné, Robert M., 6, 9, 30-51, 52, 68, 85, 139, 183, 190, 211, 311, 383, 384
- Galanter, Eugene H., 182, 215, 244
- Galton, 421, 425-426, 428, 434
- Gambling, reinforcement in, partial, 156-163
- Garnett, Maxwell, 427, 430
- Garret, H. E., 463
- Gelfand, S., 214, 225, 243, 296
- Generalization, applying, in making specific decisions, 56-58; frame of reference of, 53-54; relative validity of, about practice, 53, decision-making and, 54-55
- German, teaching of, application of development in psychology to, 189-199
- Gerstenhaber, M., 215, 244
- Gesell, A., 75, 457, 458
- Gestalt theory, 245-246, 339, 342, 345-346, 347
- Getzels, J. W., 65
- Gibb, E. G., 229
- Gibb, J., 557
- Gibb, Lorraine, 557

- Gibson, O. O., 108
 Gilbert, Thomas F., 164, 182, 199-208
 Gilinsky, A. S., 157
 Goals, 82
 Goldbeck, R. A., 218, 235
 Goldfarb, William, 332
 Goldner, R. H., 240
 Goodnow, J. J., 215, 222, 247
 Grade placement, readiness and, 70-73
 Grades, reinforcers in production of attitude change, 129-135
 Gray, 351, 352
 Grosslight, J. H., 157
 Group, acceptance, attributed social power and, 559-565; groupness of, 530-536; problem solving, 537-549, individual *vs.*, 235-238
 Growth, human, development and, 67-80, child development, educational practice and, 68-70, cognitive development, organization and, 77-80; curriculum, breadth of, 73-75, curriculum planning, child's voice in, 75-77, readiness, grade placement and, 70-73; of intelligence, 456-477, individual differences in rate of, 468-472
 Guilford, J. P., 222, 241, 245, 289, 417, 420, 435-456, 483, 601
 Gulliksen, H., 599, 600
 Guthrie, 386

 Hafner, A. J., 231
 Hagen, Elizabeth, 115
 Hall, John F., 82, 86
 Harlow, Harry F., 68, 83, 84, 85-97, 98, 99, 103, 104, 108, 200, 222, 247, 483
 Harlow, M. K., 103
 Harter, Noble, 315, 330, 331
 Hartley test, 292, 293, 294, 299, 304
 Havighurst, R. J., 486
 Hawthorne effect, 114
 Hebb, D. O., 40, 104, 483, 487, 489, 491
 Helson, H., 245
 Helson, H. B., 245
 Hilgard, E. R., 216, 217, 233
 Hillix, W. A., 218, 222
 Hints in problem solving, 234-235
 Hodgson, T., 227
 Holland, James G., 179
 Homme, Lloyd E., 179, 182
 Honzik, M. P., 459
 Horn, Ernest, 583
 Hovland, C. I., 215, 320
 Hull, C. L., 241, 246, 272, 276, 386

 Human engineering, 24
 Humphrey, G., 214, 245, 246
 Hunger drive, 91-93
 Hunter, I. M. L., 232
 Husen, T., 492
 Hypotheses, deductive, systematic development of knowledge and, 61-63; formulation of, 60-61; testable, developing, 59-60; testing, 66-67; validation of, teacher and, 58; variables to be tested, defining, 64-66
 Hypothetical mode, 258

 Imagery between speaker and listener, correspondence of, 516-518
 Incentives, 82, 87
 Individual problem solving, group *vs.*, 235-236
 Infancy, scores in, prediction from, 458-461
 Infant tests, selection, 457-458
 Inferential behavior in preschool children, 271-277
 Instinct theory, 87
 Instruction, student-centered *vs.* instructor-centered, 550-558
 Intellect, cognitive abilities, 440-443; convergent-production abilities, 447-449; divergent-thinking abilities, 444-447; evaluative abilities, 449-451; memory abilities, 443; structure of, 437-457, implications of, 452-456; three faces of, 435-456
 Intellectual potency, 259-263
 Intellectual processes, changing organization of, 461-463
 Intelligence, 416-478; adult, temporal changes in, 472-474; fifty-year curve of, 474-477; growth of, 456-477; growth rates, individual differences in, 468-472; high intellectual ability, achieving and underachieving boys of, 501-511; intellect, *see* Intellect; meaning and assessment of, 419-435, biological, 423-424, individual psychology, 425-428, observation and introspection, 421-423, physiological, 424-425, statistical, 428-429, verification by factorial techniques, 429-435
 Internal drives, 87, 90
 Internal stimulation, 89
 Inter-trial factors in learning efficiency, 42, 47-50; massed or distributed practice, 47-48; task scheduling, 48-50
 Intra-trial factors in learning efficiency,

- 41, 44-47; meaningfulness, 46-47; responses, number of, 45-46; stimuli, number of, 45
- Introspection, instructing subject to introspect, effect of, 123-125
- Irvine, R. P., 216
- Ivanov-Smolensky, A. G., 266
- Iverson, 351, 352
- Jackson, Hughlings, 425
- Jennings, H. S., 88
- Jewett, J. E., 182
- John, E. R., 233, 239, 240
- Johnson, D. M., 214, 246, 557, 594
- Jones, B. E., 230
- Jones, D. L., 320
- Judson, A. J., 214, 225, 227, 231, 243, 296
- Kaiser, Henry F., 604
- Kant, Immanuel, 423
- Kapos, E., 246
- Katona, George, 216, 346
- Katz, D., 230-231
- Kay, H., 231, 242
- Keller, Fred S., 327, 328
- Kelley, H., 557
- Kendler, Howard H., 7, 215, 218, 227, 271-277, 382, 383, 384-393, 395, 404, 436, 482
- Kendler, Tracy S., 218, 227, 271-277
- Kennedy, W. A., 99, 104
- Kersh, Bert Y., 212, 277-286, 382, 436, 551
- Kinsey, Alfred C., 90
- Kittell, J. E., 279
- Knowledge, development of, systematic, deductive hypotheses and, 61-63, theory and practice in, 63-64; as taxonomy category, 572-575
- Koffka, Kurt, 342
- Köhler, Wolfgang, 342
- Kolers, P. A., 226
- Kopstein, Felix F., 212, 311-321
- Krasnogorsky, N. D., 266
- Laboratory, classroom and, link between, 5-80; investigation of learning, relevance of, to self-instructional programming, 199-208
- Lambert, W. E., 236
- Landsman, T., 556
- Language learning, 332-334
- Lawson, 222
- Learning, associative factors, 31, 34-35, 41-50; associations, nature of, 41, 42-44, inter-trial, 42, 47-50, intra-trial, 41, 44-47; context, in communication, 330-332; by direct discovery, motivating effect of, 277-288; efficiency, criteria of, 32-33, factors in, 30-51; heuristics of discovery, 267-269; individual differences and, 23; laboratory investigation of, relevance of, to self-instructional programming, 199-208; language, 332-334; meaningful, 286; measurement of, 566-625, analysis of classroom tests, procedures for, 586-597, educational objectives and curriculum development, 569-575, educational process, in relation to, 575-586, validity of tests, 597-613, 614-625; motivation and, 23; process, important variables of, 388-390; programed, *see* Programed learning; readiness factors, 31, 34, 36-41, motivation, 36-39, reinforcement, 39-40, set, 40-41; reading, methods of, 339-355, theory of, 60-61; science of, technology of educational methods and, 21-30; by self-discovery, 285-286; situation, manipulable conditions of, 330-336, ordering of, 136-140; social, *see* Social learning; theory, 23, developments in, 390-392; transfer of, 211
- Levitt, E. E., 222
- Lewin, K., 118
- Lewis, Donald J., 156-163
- Li, J. C. R., 282
- Licht, Leonard, 287-310
- Lindquist, E. F., 313, 599
- Ling, B. C., 233, 241
- Littell, William, 569, 614-625
- Loeb, 87
- Logic, teaching and, 363-366; thinking and, 356-363
- Look-and-say reading method, 341-343
- Lorge, Irving, 229, 235, 236, 237, 323, 324, 325-338, 373, 492, 493, 512, 529, 530-536, 550, 560
- Lovell, K., 493
- Lower class, 514-515, 516-519, 520-522
- Luchins, A. S., 221
- Luria, A. L., 266
- Macfarlane, Jean W., 459
- Macomber, F. G., 383, 403-415, 528
- Madden, E. H., 214, 222, 226, 241, 246, 247
- Maier, N. R. F., 226, 271, 272, 275-276
- Maimonides, 256-257, 263

- Maltzman, Irving, 209, 212, 214, 215, 217, 221, 223, 227, 234, 241, 242, 243, 287-310, 311
- Manipulatory motivation in children, 98-106
- Marks, M. R., 219, 234-235
- Marquart, Dorothy Irene, 236, 237, 238, 529, 537-549
- Marx, M. H., 218, 222, 247
- Maslow, A. H., 93
- Massed practice, factor in learning efficiency, 47-48
- Mass media, 381-415; ETV research in decade ahead, 394-403; stimulus-response psychology, audio-visual education and, 384-393; television instruction, large and small classes, effectiveness compared, 403-415
- Maturation, 67-68, 71
- Mayzner, M. S., 221, 242
- McCall, William A., 334
- McClearn, G. E., 103
- McClelland, D. C., 104
- McCurdy, H. G., 236
- McDonald, Frederick J., 9, 52-67, 68, 85, 129, 184, 200, 384
- McDowell, 350, 351
- McKeachie, W. J., 528, 529, 550-558
- McNemar, O. W., 239, 462
- Meaningful learning, 286
- Meaningful material, verbal context and recall of, 367-380
- Meaningfulness, 46-47; validity of tests and, 609-611
- Measurement of learning, 566-625; analysis of classroom tests, procedures for, 586-597; educational objectives, curriculum development and, 569-575; educational process, in relation to, 575-586; validity of tests, 597-613, alternatives, 609-612, comment on, further, 614-625, criterion, definition of, 607-609, difficulties with, 599-602, reasons for, 603-606
- Mech, E. V., 220, 231, 246
- Mediation, 270
- Melton, Arthur W., 6-7, 9, 21-30, 31, 85, 184, 200, 356, 384, 482, 530
- Memory, abilities of intellect, 443; conservation of, 269-270
- Mendelian theory, 433
- Meyer, D. R., 92, 103
- Meyer, Susan R., 179, 182
- Middle class, 515, 517, 519, 522-523
- Miles, R. C., 103
- Mill, J. S., 427
- Miller, George A., 269, 367-380
- Montgomery, K. C., 83
- Moore, O. K., 230, 235, 236, 237
- Moos, D., 217, 234
- Moraes, A. M. de M., 240
- Morse, Samuel, 327, 328
- Mosher, 351
- Mosier, C. I., 600, 608
- Moss, G., 229, 235
- Motivation, 81-135; frustration, quality of, performance and, 117-128; grades as reinforcers in production of attitude change, 129-135; learning and, 23; learning by directed discovery, effect of, 277-280; manipulatory, in children, 98-106; motives, mice, men, monkeys and, 85-97; readiness factor of learning conditions, 36-39, aspiration, levels of, 38-39, intrinsic *vs.* extrinsic, 37-38; set applied to student teaching, 107-116
- Motives, 85-97; intrinsic and extrinsic, 263-267
- Murphy, M. J., 157
- National Defense Education Act, 399
- National Military Establishment, 25
- Nelson, 462
- Newhall, 351
- Newton, Sir Isaac, 257
- Noble, 233, 247
- Ogden, M. H., 473
- Operant, 138; conditioning, 159
- Opinion, statements of, reinforcement of, 140-156
- Order effects, 227-228
- Organization, cognitive development and, 77-80
- Originality, training of, studies in, 287-310
- Osgood, C. E., 243
- Ovsiankina effect, 286
- Pain-pleasure theory, 87
- Pavlov, Ivan Petrovich, 266
- Pepitone, A., 557
- Performance, quality of, frustration and, 117-128
- Personality, frustration and, effects of, 120-123
- Pettigrew, T. F., 222
- Phonetic training, 343-349
- Piaget, Jean, 424, 491
- Plashinski, D., 320
- Plato, 420, 421, 423, 424, 434

- Pope, Alexander, 287
 Porter, Douglas, 182
 Postman, 7
 Practice, educational, improvement of, teacher and, 52-67, child development and, 68-70, decision-making and prediction of behavior change, 55-58, evaluation of critical, procedures for, 58-67, generalizations about, relative validity of, 53; massed or distributed, 47-48; time, amount of, in problem solving, 311-321
 Preschool children, inferential behavior in, 271-277
 Presentation, pace of, as determiners of learning, 311-321
 Presssey, Sidney L., 183, 484, 578-579, 584
 Preavailability, 224-227
 Problem solving, human, 209-321; discovery, act of, 254-270; group, 537-549; inferential behavior in preschool children, 271-277; learning by directed discovery, motivating effect of, 277-288; practice time, amount of, 311-321; presentation, pace of, 311-321; research on, recent, 212-248, definitions, 214-216, difficulty, 233-234, hints and aids, 234-235, individual *vs.* group, 235-238, presenting the problem, methods of, 228-231, processes, 238-241, theory, 241-247, transfer following variations in training, 216-228, variation among elements of the problem, 231-233, variations within the problem, 228-235; training of originality, studies in, 287-310; trials, number of, 311-321
 Processes, problem solving, 238-241
 Programed learning, 136-208; content of conversation, control of, 140-156; German, teaching of, application of development in psychology to, 189-199; intrinsic programming, rationale of, 183-189; laboratory investigation of learning, relevance of, to self-instructional programming, 199-208; learning situations, ordering of, 136-140; reinforcement, in a gambling situation, partial, 156-163, of statements of opinion, 140-156; teaching machines, 164-182
 Programming, intrinsic, rationale of, 183-189; self-instructional, relevance of laboratory investigation of learning to, 199-208
 Psychological test, use of, 618-619
 Psychological theory, implication of structure of intellect for, 452
 Psychology, communication viewed by, 325-338; development in, applied to teaching of German, 189-199; education and, 1-3; individual, 425-428; of individual differences, education and, 482-500; stimulus-response, audio-visual education and, 384-393
 Raskin, David, 287-310
 Ray, W. S., 214, 218, 235, 247
 Readiness, factors of learning conditions, 31, 34, 36-41; motivation, 36-39, reinforcement, 39-40, set, 40-41; grade placement and, 70-73
 Reading, dispute about, 338-355; empirical evidence, 349-353; learning to read, methods of, 339-355, theory of early stages of, 60-61; look-and-say method, 341-343; phonetic training, for and against, 343-349
 Recall, 164, 367-380
 Recognition, 164
 Redundancy, 326, 330
 Reid, J. W., 234
 Reinforcement, 139; in gambling situation, partial, 156-163; readiness factor of learning, 39-40; of statements of opinion, 140-156; theory, 93
 Relevance, analysis of classroom tests for, 588-594
 Repetition, 326
 Research, basic and applied, 28-29; ETV, in decade ahead, 384-393; human problem solving, recent, 212-248, definitions, 214-216, difficulty, 233-234, hints and aids, 234-235, individual *vs.* group, 235-238, presenting the problem, methods of, 228-231, concreteness, 228-229, work and rest, distribution of, 229-230, processes, 238-241, theory, 241-247, transfer following variations in training, 216-228, variation among elements of the problem, 231-233, variations within the problem, 228-235; overvaluation of, 29
 Response, 138; number of responses, 45-46; similarity, 42-44
 Rest and work, distribution of, in problem solving, 229-230
 Rewards, 139
 Rhine, R. J., 221
 Ribot, 424

- Richards, 462
 Richardson, J., 247
 Richter, C. P., 88
 Riesman, David, 267
 Riley, D. A., 229-230, 233, 247
 Roby, T. B., 93, 104
 Roethlisberger, F. J., 114
 Roshal, S. M., 312, 389
 Rousseau, Jean Jacques, 68, 75
 Russell, D. H., 214, 350
- Sapon, Stanley M., 182, 189-199
 Sato, T., 220
 Saugstad, P., 214, 224, 226, 229, 245
 Schatzman, Leonard, 511-526
 Schmidt, Bernardine, 494
 Schwartz, R. J., 216, 217, 233
 Score discrimination, analysis of, 495-496
 Scott, W. A., 130
 Seashore, 343
 Segregation, 418
 Self-discovery, learning by, 285-286
 Self-instructional programming, relevance of laboratory investigation of learning to, 199-208
 Selfridge, Jennifer A., 367-380
 Sensory deprivation, 83
 Serial stimulation, 87
 Set, application to student teaching, 107-116; complex sets, 224-227; order effects, 227-228; in problem solving, 220-228; readiness factor of learning, 40-41; simple sets, 220-224
 Sex drive, 86
 Shaklee, A. B., 230
 Shannon, 345
 Shaw, Marjorie E., 237, 238, 534, 535, 538, 542, 544-545, 549
 Sheffield, F. D., 93, 104
 Sheffield, V. F., 161
 Sherburne, E. G., Jr., 382, 383, 394-403, 404
 Sherrington, 425
 Siegel, Laurence, 383, 403-415, 528
 Silva, J., 227, 246
 Silverman, R. E., 320
 Similarity, response, 42-44; in serial tasks, 44; stimulus, 42
 Simon, Seymore, 287-310
 Skinner, B. F., 7, 9, 10-20, 39, 49, 117, 137-139, 164-182, 183-184, 200, 205, 208, 254-256, 267, 301, 356, 382, 384, 416, 417, 419, 436, 527, 575, 576, 584, 585
 Skinner box, use of, 94
 Smith, B. Othanel, 339, 356-366
 Smith, H. C., 557
 Smith, W. M., 209, 221, 234
 Social class, modes of communication and, 511-526; classifications and classified relations, 518-519; differences between, 514; frameworks and stylistic devices, organizing, 519-523; imagery between speaker and listener, correspondence of, 516-518; lower class, 514-515, 516-519, 520-522; middle class, 515, 517, 519, 522-523; perspective or centering, 514-515
 Social learning, 527-565; attributed social power, group acceptance and, 559-565; group, groupness of the, 530-536; group problem solving, 537-549; instruction, student-centered *vs.* instructor-centered, 550-558
 Social power, attributed, group acceptance and, 559-565
 Solley, C. M., 232
 Solomon, H., 237, 535
 Spearman, 431, 483, 496, 498
 Spence, K. W., 7, 267, 386
 Spencer, Edmund, 423, 424
 Spiegel, J., 229, 235
 S-R theory, 91, 95
 Staats, A. W., 225
 Staats, C. K., 236, 237, 243
 Steffens, Lincoln, 267
 Stewart, J. C., 157
 Stimulation, external, 88; internal, 88; serial, 87; tactual, 89
 Stimulus, mediating, 43; number of stimuli, 45; similarity, 42
 Stolurow, L. M., 227, 246
 Stout, 427
 Strassburger, R. C., 93
 Strauss, Anselm, 511-526
 Structure of intellect, 437-451; implications of, 452-456, for education, 454-456, for psychological theory, 452, for vocational testing, 452-453
 Student teaching, set applied to, 107-116
 Süllwold, F., 240, 241
 Sully, 427
 Summers, S. S., 302
 Symonds, P. M., 479
 Székely, L., 217, 227, 241
- Tactual stimulation, 89
 Taine, 423-424

- Tapes, 381-393
- Tasks, intellectual, complex *vs.* routine, 126-127; kind of, effect of, 125-127; motor, complex *vs.* routine, 125-126; scheduling, 48-50; serial, similarity in, 44; set, 40-41; taxonomy of, 26-27
- Tate, 351
- Taxonomy category, knowledge as, 572-575; of tasks, 26-27
- Taylor, D. W., 235, 236, 240, 312
- Taylor, Harold, 585
- Teacher, improvement of educational practice and, 52-67; decision-making and prediction of behavior change, 55-58; evaluation of critical, procedures for, 58-67; student, set applied to teaching by, 107-116; validation of hypotheses and, 58
- Teaching, German, application of development in psychology to, 189-199; logic and, 363-366; methods of, 553-554; by television, 397
- Teaching machines, 164-182, 189, 200-208; easiness of material, 177-180; programming material, 168-177; questions relative to, 180-182
- Television, 381-383; educational research in, 394-403; effectiveness of, in large and small classes, 403-415
- Terman, 498
- Terrell, Glenn, 83, 98-106
- Tests, classroom, analysis of, procedures, for, 586-613; *see also* Classroom tests; definition of, operational, 623-624; psychological, use of, 618-619; reliability and standardization of, as "distractors," 624; validity of, 597-613; *see also* Validity of classroom tests
- Thinking, logic and, 357-363
- Thorndike, Edward L., 40, 41, 87, 115, 334, 335, 356, 373, 429, 464, 535
- Thurstone, L. L., 437, 464
- Thurstone, Thelma G., 583
- Toops, H. A., 603
- Training, amount of, 219-220; methods of, different, 216-219; originality, studies in, 287-310; phonetic, 343-349; transfer of, 211, following variations in, 216-228
- Transfer of training, 211; following variations in, 216-228
- Traver, R. W., 60
- Tresselt, M. E., 221, 242
- Trials, number of, as determiners of learning, 311-321
- Tropistic theory, 88
- Troyer, 584
- Tuckman, J., 229, 235
- Understanding, communication, 334-335; methods of, 216-218
- Underwood, B. J., 7, 214, 215, 222, 230, 244, 246, 247
- Validity of classroom tests, 597-613; alternatives, 609-612; appearances, mistrust of, 622-623; comments on, further, 614-625; criterion, definition of, 607-609; definitions of, 599-600; difficulties with, 599-602, reasons for, 603-606; empirical validation, need for, 619-621; essentialness of, 602; evidence of, 601; importance and convenience, 611-612; meaningfulness and, 609-611; overgeneralization of, 606-607; philosophic adequacy, 603-606; predictive, 608-609; scientific adequacy, 603; specificity of, 609; types of, 600-601
- Van de Geer, J. P., 214, 221, 223, 224, 227, 231, 241, 243-244, 245
- Verbal context, 369-377
- Vernon, Philip E., 481, 482-500, 501, 502, 512
- Verplanck, William S., 140-156, 560
- Vigotsky, L. S., 265
- Vinake, W. E., 214
- Vocational testing, implication of structure of intellect for, 452-453
- Ward, 427
- Warner, W. L., 526
- Waterhouse, Ian K., 84, 117-128, 157, 164, 483
- Watson, J. B., 86, 87, 548
- Weaver, H. E., 214, 222, 226, 241, 246, 247
- Wechsler, 476
- Weldon, 268
- Werner, 246
- Wertheimer, M., 232, 245, 342
- Whipple, J. E., 216
- White, Robert W., 263-264
- Wiesley, M., 99
- Wiggins, J. G., 222-223
- Wilson, D. P., 290
- Wissler, Clark, 428
- Wittrock, M. C., 107-116, 213

Woodring, Paul, 338-339
Work and rest, distribution of, in
 problem solving, 229-230
Wright, S., 312
Wulff, J. J., 104
Wundt, 424, 428

Yerkes-Dodson law, 93
Young, P. T., 108
Zajonc, R. B., 262
Zander, Alvin, 529, 539, 559-565
Zeigarnik effect, 286



138

